

**RIVERSIDE TEXTBOOKS
IN EDUCATION**

EDITED BY ELLWOOD P. CUBBERLEY

**PROFESSOR OF EDUCATION AND
DEAN OF THE SCHOOL OF EDUCATION
STANFORD UNIVERSITY**

RIVERSIDE TEXTBOOKS IN EDUCATION

BY THE SAME AUTHOR

MEASURING THE RESULTS OF TEACHING

**AN INTRODUCTORY BOOK ON THE USE AND VALUE OF
EDUCATIONAL TESTS, AND THEIR RELATION TO
THE WORK OF THE TEACHER**

297 pages, 45 figures in the text, 27 tables

INTRODUCTION TO THE THEORY OF EDUCATIONAL MEASUREMENTS

**A MORE ADVANCED TEXT. DESIGNED TO AID STUDENTS AND
SCHOOL OFFICERS TO MAKE CRITICAL STUDIES OF
EDUCATIONAL TESTS AND TO FORM INTELLIGENT
JUDGMENT AS TO THEIR USEFULNESS**

364 pages, 13 figures in the text, 28 tables

EDUCATIONAL TESTS AND MEASUREMENTS

BY

WALTER SCOTT MONROE, Ph.D.

PROFESSOR OF EDUCATION AND
DIRECTOR OF THE BUREAU OF EDUCATIONAL RESEARCH
UNIVERSITY OF ILLINOIS

ASSISTED BY

JAMES CLARENCE DEVOSS, A.M.

PROFESSOR OF PSYCHOLOGY AND
DEAN OF THE UPPER DIVISION, STATE TEACHERS COLLEGE
SAN JOSÉ, CALIFORNIA

AND

FREDERICK JAMES KELLY, Ph.D.

PRESIDENT OF THE UNIVERSITY
OF IDAHO, MOSCOW, IDAHO

REVISED AND ENLARGED EDITION



HOUGHTON MIFFLIN COMPANY

BOSTON · NEW YORK · CHICAGO · DALLAS · SAN FRANCISCO

The Riverside Press Cambridge

370.2
M 75 E

acc. no: 7060

COPYRIGHT, 1924, BY W. S. MONROE AND J. C. DEVOSS

COPYRIGHT, 1917, BY W. S. MONROE, J. C. DEVOSS AND F. J. KELLY

ALL RIGHTS RESERVED, INCLUDING THE RIGHT TO REPRODUCE
THIS BOOK OR PARTS THEREOF IN ANY FORM

The Riverside Press
CAMBRIDGE • MASSACHUSETTS
PRINTED IN THE U.S.A.

PREFACE TO THE REVISED EDITION

THE rapid progress which has taken place in the field of educational measurements since the first edition of this book was prepared, nearly seven years ago, has made necessary a complete rewriting of the text and the preparation of new chapters and bibliographies. In making the revision, a chapter on intelligence tests and a chapter on geography and history tests have been added. A careful examination of the book will reveal that, while much new material has been introduced, all material revised, and the size of the book materially increased, there have been, however, but few changes in the general structure of the original text. The general plan and purpose have been retained. The text is designed, as before, for beginning classes in normal schools and teachers' colleges. It is hoped that, in its new and enlarged form, the book may receive as generous a welcome as did the original edition.

Other duties have prevented Dean Kelly from having a part in this revision. Professor DeVoss has contributed the chapter on "Intelligence Tests" in addition to rewriting the one on "Handwriting." He has also read the entire manuscript.

WALTER S. MONROE

URBANA, ILLINOIS

PREFACE TO THE FIRST EDITION

THIS book is designed primarily for teachers. It is based on two years' experience in giving a course on educational measurements to prospective teachers in a state normal school and on the experience received from directing a Bureau of Educational Measurements and Standards.

It is just twenty years since Rice startled the educators of this country by his proposal that the results of teaching spelling could be measured by a spelling test. His proposal was greeted with sarcasm and ridicule, but during the past two decades the opposition to the principle of educational measurements has almost entirely disappeared. To-day the widespread use of standardized tests and scales bears witness to the importance of this movement in American education. However, it is profitable to analyze our present interest in educational measurements. A thing may be interesting merely because it is new and spectacular. Scores are objective and are subject to graphical representation. A chart displayed attracts attention. Evidence is not wanting to show that a considerable number of teachers look upon educational measurements merely as an interesting topic for teachers' meetings or as a means of attracting attention in their community.

Standardized tests and scales are not "playthings." Neither are they teaching devices. They are instruments which furnish the teacher (1) with detailed and definite aims, and (2) with a means for diagnosing the teaching situation which she faces. Unless the diagnosis is followed by remedial instruction the use of standardized tests and scales cannot be of much value. They become mere "playthings."

Our present tests are probably crude instruments, but the first railway locomotive was also crude. Even now standardized tests and scales are superior to ordinary examinations, but, more important, their use tends to engender in the teacher a type of thinking about her work which is very helpful. By using them she recognizes objective standards to be attained and not to be exceeded, the present achievements of her pupils, and that instruction must be suited to the needs of her pupils. When a teacher comes to think of her teaching problem in these terms, she is in a position to increase greatly her efficiency.

This book is addressed to teachers because they are charged with the instruction of pupils. The superintendent, principal, or student of education who is interested in the teacher's work also will find much of value in the book. Technical details of the derivation of tests are not given, but references are given so that one interested may pursue the matter. These were omitted because they are not essential to the use of the tests by teachers. For much the same reason the criticism of tests is made a secondary matter. The detailed criticism of tests and the derivation of improved ones must be left to the expert. The teacher needs to know only enough to enable her to choose wisely in selecting a test, and to prevent her from ascribing to the scores a significance which is not justified.

The newness of the field and the rapidity with which it is developing places limitations upon an attempt to write a text. It is recognized that probably before this volume is printed new tests will have been announced. However, the author believes that the point of view upon which the book is based is not merely temporary, and that, as new tests are available, the fundamental principles of the book may be applied to them.

It is obvious that in an endeavor such as this one must

utilize the results obtained by many investigators. In fact it is hoped that this book may have the virtue of summarizing these results. The author is keenly aware of his obligation to all whose work is mentioned in the following pages. Special mention should be made of Professor DeVoss, who contributed the chapter on "Handwriting," and of Dean Kelly, who wrote the chapter on "Reading."

WALTER S. MONROE

EDITOR'S INTRODUCTION

UP to very recently our chief method for determining the efficiency of a school system was the method of personal opinion. When the work of a superintendent of schools was called in question, the schools were visited and personal opinions expressed as to their standing. In case of a disagreement among the visitors the efficiency became a matter of dispute, and the people of a community usually favored the opinion which most nearly coincided with their prejudices and preconceived ideas.

Relatively recently the method of comparison was introduced. By means of this method the school system under consideration is compared with other school systems of the same size and class, and with reference to a number of different items. After such a comparison has been made, it is possible to place the school system relatively. If the school system studied stands fourth out of twenty school systems compared in one item, thirteenth in another, and at the bottom of the list in three others, it is not difficult to determine its position. It is evident that this is a much better method than the one of personal opinion. Its chief defect, though, lies in that the school system studied is continually compared with the average or median of its size and class. In other words, the school system is continually measured as against mediocrity, when as a matter of fact the average or median school system may not represent a good school system at all. Perhaps all of the school systems below the average or median should be classed as poor school systems, and even some of those above are not doing what a school system should do.

Still more recently, and wholly within the past decade, a still better method for the evaluation of the work which teachers and schools are doing has been evolved. This new method consists in the setting up, through the medium of a series of carefully devised "Standardized Tests," of standard measurements and units of accomplishments for the determination of the kind and the amount of work which a school or a school system is doing. This new movement is as yet still young, but so important has it become in terms of the future of school administration that it already bids fair to change, in the course of time, the whole character of this professional service.

The significance of these new standards of measurement for our educational service is indeed large. Their use means nothing less than the ultimate transformation of school work from guesswork to scientific accuracy; the elimination of favoritism and politics from the work; the ending forever of the day when a personal or a political enemy of a superintendent can secure his removal, without regard to the efficiency of the school system he has built up; the substitution of well-trained experts as superintendents of schools for the old successful practitioners; and the changing of school supervision from a temporary or a political job, for which little or no preparation need be made, to that of a highly skilled piece of social engineering.

This new method for the evaluation of the work which a school system is doing is so important that any young man or woman of to-day who desires to prepare for school administration should by all means thoroughly familiarize himself or herself with the aims and methods of this new type of administrative procedure. The underlying purpose of the new movement has been the creation of such standardized scales for measuring school work, and for comparing the accomplishments of different schools and groups of school children,

as to give to both supervisors and teachers definite aims in the imparting of instruction. Instead of continuing to teach without definite measuring-sticks, and to assign tasks and trust to luck and the growth process in children for results, which is comparable to the old-time luck-and-chance farming, it has been attempted to evolve standards of measurement which will do for education what has been done for agriculture as a result of the application of scientific knowledge and scientific methods to farming.

Such an important new movement is of especial significance to the teacher in charge of a class, to the citizen interested in schools, and to the superintendent responsible for results.

To the teacher it cannot help but eventually mean not only concise and definite statements as to what she is expected to do in the different subjects of the course of study, but the reduction of instruction to those items which can be proved to be of importance in preparation for intelligent living and future usefulness in life. It will mean, too, an ultimate differentiation in training for the different types of children with which teachers now have to deal, and the specialization of work so as to enable teachers to obtain more satisfactory individual results. To the citizen the movement means the erection of standards of accomplishment which are definite, and by means of which he can judge for himself as to the efficiency of the schools he helps to support. For the superintendent it means the changing of school supervision from guesswork to scientific accuracy, and the establishment of standards of work by which he may defend what he is doing.

Up to the present time nearly all of the work which has been done in the evolution and testing out of these new standardized tests has been work of a highly scientific and technical nature, most of the articles being written in a

language which the layman can scarcely understand. Often no interpretation has been attempted of the results which have been obtained. The classroom teacher and the school principal have naturally not found these studies of much help to them in their work.

This work has been carried far enough, however, so that the time now seems ripe for a clear and simple statement as to the nature of the different tests which have been evolved, their use, their reliability, what are the best standard scores so far arrived at, and, in particular, how to diagnose the results and apply remedial instruction. This the three authors of the present volume in the series have attempted to give, and, to make their work of the largest possible usefulness to normal-school students, teachers, and principals of schools, they have cast the whole in language so simple and untechnical that the average grade teacher can read the book and understand it. In addition, to give still larger value to the book, they have added a number of chapters, written in a similar simple and readable style, dealing with the construction of tests, testing programs, the meaning of scores, the improvement of examinations, and the use of the standardized tests in the work of school supervision.

No space has been taken up in merely reproducing the tests themselves, though samples, showing their nature, have been inserted. If it is desired to use the tests with a class, they will be needed in quantities, and they may then be obtained in quantities and for very small sums from the persons and at the places mentioned in Appendix B. The chapter bibliographies give the most important book or article describing in detail the construction and use of the tests, in case the worker desires to go further than this volume presents. Instead, the authors have used their space in explaining to teachers and school officers the nature of the tests, telling how to give and score them, what standings

the pupils should attain in their use, and presenting a rather full description as to the significance of the results obtained and how to remedy the defective conditions which the use of the tests reveals. In consequence, the book should prove of much use not only to students in normal schools and colleges, but to teachers and principals in our public schools as well.

ELLWOOD P. CUBBERLEY

CONTENTS

CHAPTER I. INTRODUCTION 1

Measuring children's abilities — Teachers' estimates and examination grades — Recent investigations have revealed inaccuracy — High grades and low grades — Errors in marking papers — Marks imply subjective norms — Limitations in content of ordinary examinations — Two remedies — Standardized tests improved measuring instruments — Not perfect — Intelligent attitude toward — Plan of the following chapters — How to study them — Questions and topics for investigation.

CHAPTER II. ARITHMETIC 19

- I. The problem of measuring arithmetical abilities — Purpose in measuring — Arithmetical abilities specific — Types of examples — Characteristics of arithmetical abilities — Of the ability to solve problems — General *vs.* diagnostic tests.
- II. Standardized tests for measuring abilities in the operations of arithmetic.
 1. The Curtis Standard Research Tests, Series B.
 2. The Curtis Standard Supervisory Tests in Arithmetic.
 3. The Cleveland-Survey Arithmetic Tests.
 4. The Lunceford Diagnostic Tests in Addition.
 5. Monroe's Diagnostic Tests in Arithmetic.
 6. Monroe's General Survey Scales in Arithmetic.
 7. The Woody Arithmetic Scales.
 8. Woody-McCall Mixed Fundamentals.
 9. Peet-Dearborn Progress Tests in Arithmetic.
 10. Lippincott-Chapman Arithmetic Fundamentals Test.
General structure — Use of — Function — Norms — Limitations of each.
- III. Standardized tests for measuring the ability to solve problems — The process of problem-solving — Requirements for a reasoning test — Function of.
 1. Monroe's Standardized Reasoning Tests in Arithmetic.
 2. Buckingham's Scale for Problems in Arithmetic.
 3. Other reasoning tests —
Stone — Curtis — Lippincott-Chapman.
- IV. Interpretation of scores and remedial instruction — Accuracy of individual scores — Gain or loss in repetition — Accuracy in terms of coefficients of reliability — Accuracy of class scores — Intelligence and achievement — Scientific

management — Diagnosis of a teaching situation — A plan of diagnosis — Typical cases — Meeting the situation; laws — Individual *vs.* class needs — Modifying the class drill — Use of practice tests — Questions and topics for investigation — Selected bibliography.

✓ CHAPTER III. READING 94

I. Silent Reading — Complex nature of silent reading — Traits to be measured — Types of reading — The problem of measurement — General limitations of silent reading tests.

1. Monroe's Standardized Silent Reading Tests.
2. Monroe's Standardized Silent Reading Tests, Revised.
3. Courtis Silent Reading Test No. 2.
4. Burgess Picture Supplement Scale.

General Structure — Giving — Functions — Norms — Limitations, for each.

5. Thorndike's Scale Alpha for Measuring Understanding of Sentences.
6. Same, Scale Alpha 2.
7. Haggerty's Achievement Examination in Reading, Sigma 1.
8. Same, Sigma 3.
9. Thorndike-McCall Reading Scale.

All require answers to questions with text in hand — Nature of pupil's performance — Function — Scores — Norms — Limitations of these scales.

II. Vocabulary Scales — The problem of measurement — Facts of title — Nature of pupil's performance — Selection of test words — Pupil's performance — Function of tests — Limitations.

III. Gray Oral Reading Test.

General structure — Giving — Pupil's performance — Proposed modification — Norms — Limitations.

IV. Interpreting scores and planning remedial instruction — Limitation to be kept in mind — Interpreting scores of classes — Raising median comprehension score — Care — Reading above primary grades — Reading in upper grades — Questions and topics for investigation — Selected bibliography.

CHAPTER IV. HANDWRITING 155

I. The problem of measurement of handwriting — Traits measured.

II. Measuring the act of handwriting — Emphasis on position — Measuring movement — Results.

- III. Measurement of rate of handwriting — General procedure — Time — Copy — Directions to pupils.
- IV. Measurement of quality in handwriting — Thorndike's Handwriting Scale — Ayres Handwriting Scale — Starch Handwriting Scale — New York City Penmanship Scale — Freeman Handwriting Scale — The score card for detailed analysis of — Objective indices of legibility — Methods of using handwriting scales — Measurement for diagnosis — Use of score card — Using the Freeman Handwriting Scale.
- V. The reliability of measures of handwriting — Reliability as to rate — Reliability as to quality — Scales compared — Objectiveness of scores — Training in use of scale.
- VI. Norms — Two types of norms — Norms of progress — Norms of attainment — Norms of attainment for maximum rate and quality.
- VII. Procedure after measurement — Extremes to be avoided — Interpreting the scores of a class — Using scores to motivate drill — Self-measurement — Directing the practice drills — Systems of penmanship — Rhythm — Speed — Reasons for using scales — Questions and topics for investigation — Selected bibliography.

✓ CHAPTER V. SPELLING 205

- I. The problem of measurement in spelling — Difficulties encountered.
- II. Standardized Lists of Foundation Words.
 - 1. The Ayres Spelling Scale.
 - 2. The Iowa Spelling Scales.
 - 3. Second and third thousand most frequently used words.
- III. Spelling tests — Selecting words for a uniform test — How difficult words to use — Selecting words for a scaled test — How many words to use — Methods of giving the test — Letters per minute — Summary — Directions for giving a timed sentence test — Grade norms.
 - 1. Monroe Timed Sentence Spelling Tests.
 - 2. Courtis Standard Dictation Spelling Tests.
 - 3. The Seven S Spelling Scales.
 - 4. Starch's Spelling Scale.

Structure and norms for each.
- IV. Interpretation of scores and remedial instruction — General interpretation — Individuality in spelling difficulty — Types of spelling difficulties — Jones's Spelling Demons — Teaching pupils to correct their errors — Causes of some misspellings — Good teaching of spelling — Devices for improving — Making associations automatic — Courtis's spelling

practice tests — Questions and topics for investigation —
Selected bibliography.

✓ CHAPTER VI. ENGLISH 240

I. The problem of measurement in English — The nature of ability in language — In composition — In literature.

II. Language and grammar tests.

1. Charters Diagnostic Language Tests.
2. Charters Diagnostic Language and Grammar Tests.
3. Briggs English Form Test.
4. Kirby Grammar Test.
5. Pressey Diagnostic Tests in English Composition.
6. Starch Grammatical Scales.
7. Starch Punctuation Scale.
8. Wilson Language Error Test.
9. The Boston Copying Test.
Structure — Use — Norms.

III. English composition.

1. Directions for securing compositions — Points to be covered in directions — Directions for using the Hillegas Scale.
2. Composition scales — General structure — The Hillegas Scale — Nassau County Supplement — Hudelson Scales — Van Wagenen English Composition Scales — Courtis Standard Research Test in Composition — Lewis Composition Scales.
3. Using composition scales — Plan — Grade norms — Reliability — Reliability when corrected for mechanical errors — Hudelson's index of reliability.

IV. Literature — The problem of measurement — Abbott and Trabue Scale.

V. Educational significance of the use of these tests and scales — Finding specific language weaknesses — Remedying the situation revealed — Questions and topics for investigation — Selected bibliography.

CHAPTER VII. GEOGRAPHY AND HISTORY 272

I. Geography — The problem of measurement in geography.

1. Hahn-Lackey Geography Scale.
2. Courtis Supervisory Tests in Geography.
3. Gregory-Spencer Geography Tests.
4. Posey-Van Wagenen Geography Scales.
5. Witham Standard Geography Scales.
Structure — Uses.

- ✓ **II. History** — The problem of measurement in history.
1. Hahn Scale for Measuring Ability in History.
 2. Harlan Test of Information in American History.
 3. Van Wagenen American History Scales.
 4. Barr Diagnostic Tests in American History.
 5. Gregory Tests in American History.
- Structure, and criticism of these scales.
- Questions and topics for investigation — Selected bibliography.

CHAPTER VIII. HIGH-SCHOOL TESTS 297

Limitations of achievement tests for use in the high school —
Prognostic tests most valuable for use in the high school — Little
opportunity for diagnosis of high-school students as to achieve-
ment — Purposes to be realized from use of tests in high schools.

I. Mathematics — Problem of measurement in algebra.

1. Douglass Standard Diagnostic Tests for First-Year Algebra.
2. Holz Algebra Scales.
3. Illinois Standardized Algebra Tests.
4. Rugg and Clark Tests in First-Year Algebra.
5. Rogers Tests of Mathematical Ability.
6. Minnick Geometry Tests.

Structure of these tests — Meeting the teaching situ-
ation revealed by algebra tests.

II. Latin — Problem of measurement similar to that of English.

1. Godsey Diagnostic Latin Composition Test.
2. Henmon Latin Tests.
3. Holtz-Godsey Latin Teaching Tests.
4. Pressey Test in Latin Syntax.
5. Tyler-Pressey Test in Latin Verb Forms.
6. Starch-Waters Latin Tests.
7. Ullman-Kirby Latin Comprehension Test.

Structure of these tests.

III. Modern Languages — Problem of measurement in.

1. Handschin Silent Reading Tests in Spanish.
2. Handschin Silent Reading Tests for French.
3. Handschin Comprehension and Grammar Test A, French.
4. Henmon French Tests.
5. Wilkins Prognostic Test in Modern Languages.

Structure of these tests.

IV. Science — problem of measurement in.

1. Downing Range-of-Information Test in Science.
2. Starch Physics Test.

3. Iowa Physics Test.
4. Van Wagenen Reading Scale for General Science.
Structure of these tests.
Questions and topics for investigation — Selected bibliography.

CHAPTER IX. INTELLIGENCE TESTS 332

- I. The problem of measurement of general intelligence — Measurement of general mental ability not new — Definition of intelligence — An eclectic view — General plan of measurement.
- II. Descriptions of representative intelligence tests.
 1. Tests of the Binet-Simon type — The Stanford Revision — Scores secured — Advantages and limitations of the Stanford-Binet Scale — Reliability and validity — Retests and constancy of the I.Q. — Goddard's Revision of the Binet Scale — Kuhlman's Revision — Advantages and limitations of Kuhlman's Revision.
 2. Individual performance tests — The Army Performance Scale.
 3. Group tests for literate school children — Definition of terms — Historical — Army Alpha Group Intelligence Test — The National Intelligence Tests — Types of tests in group intelligence examinations of the Army Alpha type — Sixteen types — Other group intelligence examinations for literate school children — Limitations and advantages — General structure of.
 4. Non-verbal intelligence tests — The Army Beta Group Intelligence Examination.
 5. Group tests for college students.
- III. Using measurements of general intelligence.
Questions and topics for investigation — Selected bibliography.

CHAPTER X. TESTING PROGRAMS 377

A testing program for a general survey of achievement — Should compare intelligence and achievement — Combined scores with two or more tests — Derived scores — The Illinois Examination — Pintner Educational Survey Tests — Stanford Achievement Test — Lippincott-Chapman Classroom Products Survey Tests — Pressey's Scale of Attainment, No. 1 — Same, No. 3 — Same, No. 2 — Selecting tests for a testing program — Cost of a testing program — Questions and topics for investigation — Selected bibliography.

CHAPTER XI. THE CONSTRUCTION OF STANDARDIZED TESTS . 393

Value of a general knowledge of test construction.

- I. Principles of test construction — Determination of minimum essentials of a subject — Types of test to be constructed — Power test, rate test, or quality scale — Types of exercises — Selection of exercises for final form of test — Directions for administering — Standardization of the test — Derived scores — Critical evaluation.
- II. An illustration of the critical evaluation of a test — Function — Validity — Objectivity — Reliability — Correlation — Reliability coefficients — Discrimination — Probable error — Comparison with criterion measures — Inferences concerning validity based on structure of test and its administration — Summary for validity — Practice effect when a test is repeated.

Questions and topics for investigation — Selected bibliography.

CHAPTER XII. THE MEANING OF SCORES 417

Two methods of describing achievement — Need for interpreting scores into school marks — A satisfactory standard — An efficient standard — Effort on tool subjects — School demands on these — Basis for standards of accomplishments — Translating the scores into school marks — Use of the normal probability curve — A standard distribution as an item in school policy — Questions and topics for investigation — Selected bibliography.

✓ CHAPTER XIII. ADMINISTRATIVE AND SUPERVISORY USES OF EDUCATIONAL TESTS 431

The supervisor's responsibility — Value of a test depends on use made of measures — Administrative and supervisory uses of standardized educational tests — Two fundamental questions — Test scores not the only basis of action — Basis for determining the value of a given procedure — Two special uses.

- I. The promotion and classification of pupils — The organization of our school systems — Need for further adjustment to capacity — Flexible plans for promotion — Terman's five-track plan — Establishment of a multiple-track system not dependent on use of educational tests — Is a multiple-track plan desirable? — Arguments for — Arguments against — Determination of merits of multiple-track plan by scientific experimentation — Statements by other investigators — Accuracy of classification on basis of test scores — The ultimate effect of segregation on society — On the school —

Value of complete segregation problematical — Partial segregation recommended — Other plans for adjusting the school to the pupil.

II. Educational and vocational guidance of pupils — Meanings of — Need for — Large percentage of failures in high school an index of ineffective guidance — Educational guidance should reduce failures, but will not eliminate — Guidance in business — Educational guidance policies.

III. Information needed in educational and vocational guidance: (1) Pupil's capacity to learn; (2) pupil's interests; (3) vocational information; (4) relation of degrees of capacity to probable success — Relation between general intelligence and school success — And occupational success — Difficulties encountered in guidance work — Administration of educational guidance.

Questions and topics for investigation — Selected bibliography.

✓ CHAPTER XIV. IMPROVEMENT OF WRITTEN EXAMINATIONS . 469

Reliability of written examinations — Coefficients of reliability of standardized educational tests — Relative reliability of examinations and tests — Absolute reliability of examination grades — Constant errors in examination scores — Improvement of written examinations — Importance of the problem of measurement — Methods for increasing the reliability of written examinations — Advantages of the "new examination" — Limitations of — Decreasing the constant errors — Increasing the objectivity — Suggested rules for the administration of examinations and the marking of papers.

APPENDICES

A. GLOSSARY OF TECHNICAL TERMS 487

B. LIST OF STANDARDIZED TESTS DESCRIBED 503

INDEX 519

LIST OF FIGURES

1. Distribution of Marks Assigned to one Geometry Paper .	6
2. Showing Form of Tabulation Sheet for Recording Scores Obtained by Using the Courtis Standard Research Tests, Series B, and the Scores of a Seventh-Grade Class in Addition	31
3. Chart, Showing the Scores Made by a Sixth-Grade Pupil, in Comparison with the Standard Scores, Using the Courtis Arithmetic Tests	81
4. Showing Method of Estimating Corrections to be Added to Scores of Test II	105
5. Showing Forms Used in Recording the Scores Obtained by Using Courtis Silent Reading Test No. 2	110
6. Standard Distribution of Scores for Burgess Picture Supplement Scale	115
7. A Section of the Thorndike Handwriting Scale	165
8. Two Sections from the Ayres Handwriting Scale, "Gettysburg Edition"	167
9. Standard Score Card for Measuring Handwriting	173
10. Individual Record Card, Freeman Scale	178
11. Graphical Representation of Grade Norms Given by Ayres for the "Gettysburg Edition" of his Scale	191
12. Sixth-Grade Scores for Quality of Handwriting (Ayres Scale)	194
13. Sixth-Grade Scores for Rate of Handwriting (Ayres Scale)	195
14. Sixth-Grade Scores for Rate and Quality of Handwriting	196
15. Measuring Scale for Ability in Spelling (Ayres)	207
16. Showing the Distribution of 91 Pupils According to the Number of Words Spelled correctly	210
17. Preliminary Test of the Courtis Supervisory Test in Geography	276
18. Correlation of Form 1 Scores with Form 2 Scores of the Illinois General Intelligence Scale, Fifth Grade	406

LIST OF TABLES

I. Grade Norms for Courtis Standard Research Tests, Series B	33
II. Grade Norms for the Cleveland-Survey Arithmetic Tests	39
III. Monroe's Diagnostic Tests in Arithmetic, Grade Medians	44
IV. Showing Tabulation of Scores Made by the Individual Pupils of a Sixth-Grade Class on the Woody Addition Scale, Series A	53
V. Grade Norms for Woody Arithmetic Scales in Terms of Number of Examples done correctly	56
VI. Monroe's Standardized Reasoning Tests in Arithmetic, Form I; Grade Norms	63
VII. Buckingham's Scale for Problems in Arithmetic, Form I; Grade Norms	65
VIII. Buckingham's Scale for Problems in Arithmetic, Form I; Grade Distributions	66
IX. Grade Norms for Stone Reasoning Test	68
X. Showing Gain of Second Score over First, Courtis Standard Research Tests, Series B	71
XI. Showing Distribution of the Pupils of a City According to the Number of Examples Attempted, Courtis Standard Research Tests, Series B	79
XII. Norms for Monroe's Standardized Silent Reading Tests, Revised	107
XIII. Credit Corresponding to Each Number of Paragraphs Marked correctly in Each Grade, Burgess Picture Supplement Scale	115
XIV. Norms for Haggerty Sigma 1 and Sigma 3, and the Thorndike-McCall Reading Scales	126
XV. Credits to be Given for Complete and Partial Success in Reading a Paragraph, Gray Oral Reading Test	137
XVI. Form to be Used in Calculating Individual Scores	139
XVII. Norms for Gray's Oral Reading Test	143
XVIII. Norms of Progress Proposed by Freeman	185

XIX.	Norms of Progress for the Ayres "Gettysburg Edition"	185
XX.	Median Handwriting Scores for Rate	186
XXI.	Median Handwriting Scores for Quality	187
XXII.	Median Scores (Words Spelled correctly) and Ayres's Norms	221
XXIII.	Misspellings of 80 Seventh-Grade Pupils in a Column Spelling Test	229
XXIV.	Grade Norms for Charters's Diagnostic Language Tests	244
XXV.	Classification of Errors Made by 275 First-Year Pupils on the Monroe Standard Research Tests (Algebra)	312
XXVI.	Correlation (r) between Stanford-Binet and the Tests of the Army Performance Scale	350
XXVII.	Showing the Location of the Various Tests in Six Intelligence Examinations	361
XXVIII.	Reliability Coefficients	409
XXIX.	Probable Errors of Measurement and Ratio of Probable Errors of Measurement to Average Scores	410
XXX.	Correlation between Scores Yielded by the Illinois General Intelligence Scale and by the National Intelligence Scale	412
XXXI.	Gains due to Special Instruction upon the Illinois Examination	415
XXXII.	Summary Distribution of Coefficients of Reliability for Written Examinations	471
XXXIII.	Reliability Coefficients of Standardized Educational Tests	472

EDUCATIONAL TESTS AND MEASUREMENTS

• • •

CHAPTER I

INTRODUCTION

Traditional methods of measuring abilities of school children. In the past two methods have been employed for measuring the abilities of school children. On the basis of their general observations teachers have estimated the capacity of children to do the work of the school. Some have been classed as "bright," others as "stupid" or "dull," and still others as "just average." Estimates of achievements in school subjects have been expressed by means of grades, such as "A," "B," etc., or as 75, 82, 97, etc. At the end of the term, and in many schools at other times, more formal measurements of achievements have been secured by means of written examinations. These two methods of measurement, teachers' estimates and written examinations, have been employed since schools began. They are still and probably always will be extensively used.

Teachers' estimates and examination grades considered important. Until recently the measures which written examinations yield have been considered to possess a high degree of precision and accuracy. Much the same attitude has prevailed with reference to estimates of achievement, at least by the teacher who has made them. Both types of measures have been treated very seriously. The promotion of pupils depended upon the "grades" they received. The

2 EDUCATIONAL TESTS AND MEASUREMENTS

ability of a pupil in each of the subjects has been measured by the teacher's estimate and by written examination, and, if the resulting measures showed the pupil to be a few points, or in some instances a fraction of a point, below the "passing mark," the pupil was classified as a failure. If the resulting measures equaled or were above the "passing mark," the pupil was promoted.

The "grades" or school marks are entered upon the monthly or quarterly report cards. Parents, as well as teachers and pupils, take these school marks very seriously. If Johnnie's "grades" for a given month are below those of the preceding months, or, worse still, if they are below those of neighbor Smith's Mary, an explanation is demanded. A permanent record is kept of at least the yearly "grades," and the awarding of school honors is based upon them.

Until recently practically all admission to college was determined by examination. Except in the universities and colleges of the Central and Western American States the custom still maintains generally throughout the world. This practice is based on the assumption that the examining committee can determine thereby the effectiveness of the candidate's college preparatory work. The civil service, from its inception in China centuries ago until the present day, has employed the examination as a means for measuring the ability of persons who desire positions operated under this system.

Recent investigations have shown traditional school marks to be inaccurate. Within the last few years a number of investigations have been made to ascertain the accuracy of measures of school achievement obtained by means of teachers' estimates and by means of written examinations. In the world of physical things we measure distance by means of the yardstick, mass by means of scales, and the volume of liquids by means of a gallon measure. When two

persons measure the length of the same room by means of the same yardstick or different yardsticks the two measurements will generally be approximately equal. If they differ by more than one or two inches, we doubt the accuracy of both, and we demand that the room be measured again. Repetition of the process of measurement will generally reveal the presence of errors if the original measures are not accurate.

Much the same method has been employed in investigating the accuracy of teachers' estimates and the "grades" yielded by written examinations. If the same children are measured in the same school subject by two teachers and the two sets of measures do not agree rather closely, we have reason to doubt the accuracy of both sets of school marks. Our plan of school work does not permit "final grades" in a given subject to be assigned independently by two different teachers at the same time. For this reason it has been necessary to make comparisons between "final grades" given to the same pupils for successive terms. Although crude this method will usually reveal the presence of errors in the "final grades." Two teachers working independently may administer written examinations to the same pupils, but this method has not usually been followed in studying the accuracy of examination marks. With few exceptions the investigations of written examinations have been limited to the marking of the same paper by different teachers. Investigations of both types of measures have been unanimous in showing that our traditional school marks are inaccurate. A few of the most significant investigations will be briefly reviewed.

Evidence that "high grades" are given by some teachers and "low grades" by others. It is a well-known fact that some teachers give a much larger per cent of "high grades" than other teachers do. Occasionally a teacher or even an entire department has become notorious because of the

4 EDUCATIONAL TESTS AND MEASUREMENTS

"high grades" or "low grades" which they gave to their students. The writer has learned of one institution in which it was customary to discount the "high grades" given by one department in awarding certain school honors. In another institution a department was found which was giving a grade of failure to nearly half of the students enrolled in certain elementary courses. One is justified in being suspicious of the accuracy of the grades given by a teacher when they tend to be either very high or very low. Sometimes a teacher exhibits considerable pride because all members of his class have made very "high grades." In some cases it is doubtless true that such "high grades" are the result of superior achievements, but frequently they are high because the teacher tends to give "high grades." Many schools have adopted the policy of summarizing in a distribution the grades given by different teachers and by different departments. Even when the number of grades involved is so large as to make the summary representative, marked differences are usually found between different teachers and departments. The error introduced by the tendency to give "high grades" or "low grades" may be described as constant. It is the same type of error as occurs when a merchant gives "short weights."

A typical investigation for revealing the presence of constant errors in grades has been reported by Johnson,¹ Principal of the University High School of the University of Chicago. In the University High School, "F" denotes failure, and the four successive ranks above failure are indicated by "D," "C," "B," and "A." For the several departments of the school, Johnson tabulated the number of times each mark was given during the years 1907-08 and 1908-09. The facts revealed by these tabulations may be illustrated by the

¹ Johnson, F. W., "A Study of High School Grades"; in *School Review*, vol. 19 (January, 1911), pp. 13-24.

following. In English the per cent of failures was 15.5, which was nearly double that of history (8.1). The highest mark ("A") was given to 9.3 per cent of the pupils taking French, but in the German department it was awarded to 17.1 per cent. English and history occupied similar places in the program of studies. They were taken by practically all students. French and German likewise occupied similar places in the school. Comparisons between the distributions of the grades given by individual teachers revealed much greater differences. It was evident that some teachers tended to give "low grades" and that others tended to give "high grades." This means that such grades involved a constant error.

Evidence of errors in marking examination papers. Scientific investigation has proved that the marking of examination papers is subjective; that is, different teachers, when working independently, tend to assign widely varying marks to the same paper. An investigation by Starch and Elliot¹ is typical of many that have been made. These investigators selected a final examination paper in geometry, written by a student in one of the largest high schools in Wisconsin. An exact reproduction of this paper and a set of the questions were sent to one hundred and eighty high schools in the North Central Association. It was requested that this paper be graded according to the practice and standards of the school by the principal teacher of mathematics. One hundred and sixteen acceptable replies were received. The papers showed evidence of having been marked with unusual care and attention. The distribution of the marks is shown in Fig. 1. Of the one hundred and sixteen marks, two were above 90, while one was below 30. Twenty were 80 or above, while twenty other marks were below 60. Forty-seven

¹ Starch, Daniel, and Elliot, E. C., "Reliability of Grading High School Work in Mathematics"; in *School Review*, vol. 21 (1913), pp. 254-59.

6 EDUCATIONAL TESTS AND MEASUREMENTS

teachers assigned a mark passing or above, while sixty-nine teachers thought the paper not worthy of a passing mark.

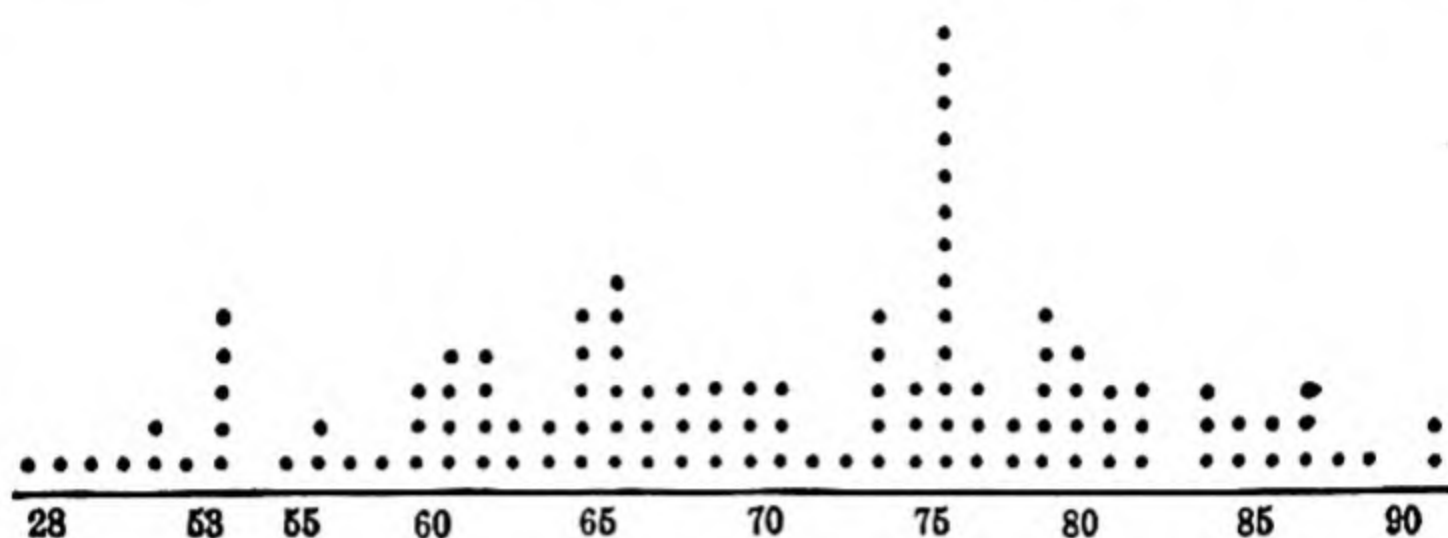


FIG. 1. DISTRIBUTION OF MARKS ASSIGNED TO ONE GEOMETRY PAPER BY 116 TEACHERS

Passing grade 75. Range 28 to 92. Marks assigned by schools whose passing grade was 70 were weighted by 3 points.

A striking illustration of the subjectivity of the marking of examination papers by college instructors is cited by a recent writer.¹ One of the group of expert readers assigned to the marking of a set of examination papers in history, after scoring a few of them, wrote out for his own convenience what he considered model answers to the questions. By some mischance this "model" examination paper fell into the hands of another expert reader who graded it as a paper written by a student. The mark he assigned to it was below passing, and, in accordance with the custom, this "model" was rated by a number of other expert readers in order to insure that it was properly marked. The marks assigned to it by these readers varied from 40 to 90.

The extreme marks assigned to an examination paper in such investigations are due in part to a constant error; that is, a tendency to give "high grades" or "low grades." The natural tendency of competent persons to differ in the judgments which they express also contributes to the variability

¹ Wood, Ben D., "Measurement of College Work"; in *Educational Administration and Supervision*, vol. 7 (September, 1921), pp. 301-34.

of the marks assigned to the same examination paper. The error resulting from this cause is *variable*. In the case of a group of examination papers marked by a person whose grades were not subject to a constant error, some papers would be marked "high," others "low." Evidence of this fact could be secured by having a set of papers marked independently by two persons. If the average of the marks assigned by one person is not equal to the average of the other set, a correction should be made by adding the amount of the difference to each mark of the set having the low average. After this has been done, it will be found that the two marks for the same paper do not agree in most cases. About half of the differences will be positive and approximately the same number negative. These differences do not represent the magnitude of the variable errors, but they are evidence of the presence of such errors.

Marks assigned to examination papers imply subjective norms. In reading the controversial literature on written examinations one will find little mention of the criticism expressed by the heading of this paragraph, but in the illustrations given below the reader will recognize that this weakness of examination marks has been sensed by most pupils and teachers. However, they appear to have failed to analyze the situation sufficiently to grasp the source of the difficulty. In the judgment of the writer this is one of the most serious weaknesses of the traditional examination.

In order to understand how norms (standards) are used in connection with the grading of examination papers, it is necessary to distinguish between scores, or measures, and "grades," or marks. A score simply describes the performance which has been recorded in the examination paper. For example, a pupil may answer 55 per cent of the questions correctly. In this case 55 is his score. If a certain number of points or credits had been given for each question his score

8 EDUCATIONAL TESTS AND MEASUREMENTS

might be 129 or 91, or 217. A "grade" interprets this description with reference to certain norms. A "grade" indirectly describes a pupil's performance on an examination, but it also tells whether the pupil's performance is to be considered as above passing or below passing; whether he is to receive the highest mark or the lowest mark or an average mark. It is customary to describe the quality of examination papers in terms of the per cent of questions answered correctly. For example, if an examination includes ten questions and a pupil answers seven of them correctly and an eighth one partially right, he is given a score of 75 per cent, which is interpreted to mean that in the judgment of the examiner he has answered the questions 75 per cent correctly. School marks or "grades" are also frequently expressed in terms of per cents. Sometimes they are expressed in terms of letters or other symbols, but these in turn are generally defined in terms of per cents. For example, the grade of "A" may be defined as being between 95 per cent and 100 per cent.

Since both scores and "grades" are generally expressed in terms of per cents, it is only natural that the two have been confused and that scores have been used as "grades." A good illustration of their difference came to the writer recently. An examination in mathematics was given to nearly one thousand freshmen in one of our large universities. This examination may properly be described as "hard," considering the training which the students had received. One student made a score of 100. The lowest score was 12. The average was approximately 55. From the standpoint of the distribution of scores this was a "good examination." If it had been easier, so that any considerable number of pupils received scores of 100 per cent, it would have been defective. If it had been so "hard" that a considerable number of students made zero scores, it would also have been defective. In

both cases it would have failed to differentiate between some students who were not equal in ability. If a passing mark of 70 or 75 had been adopted, more than three fourths of this group of students would have received a "grade" of failure. The grades as a whole would have been properly described as "low." They would have involved a constant error. The passing mark for this particular examination probably should be in the neighborhood of 40. Then a score of 40 would be translated into a "grade" of 70 or whatever passing mark this institution has adopted.

The recognition of this distinction between scores and "grades" enables us to indicate the way in which subjective norms are implied in "grades." A "grade" is not a pure measure or description of the pupil's performance. It is rather an interpretation of the measure of his performance with reference to certain norms. When no distinction is made and scores are used as "grades," pupils will receive high "grades" if the examination is "easy"; if it is "hard" they will receive low ones. Thus, the difficulty of the examination is one factor in establishing the norms with reference to which the scores are interpreted when they are used as "grades." Severe marking will tend to set high norms. It is only when the examination is of average or "standard" difficulty and the marking is average in severity that scores and "grades" should be treated as identical in magnitude. Since the norms are established by the difficulty of the examination and the severity of the scoring, they must be subjective. In the investigations of the marking of examination papers, it was shown that teachers varied widely in their judgments concerning the worth of examination papers. There is no reason to expect that they would agree more closely in estimating the difficulty of examinations. Hence, norms which depend upon the difficulty of a set of questions that represents a teacher's opinion of a "fair exami-

nation" and upon the severity of his marking of the papers must be considered subjective.

Limitations in the content of ordinary examinations. The criticism is frequently made that teachers, in formulating examination questions, tend to ask for unimportant details and to neglect the minimum essentials of a subject, and that, therefore, a pupil's performance on an examination cannot be a truthful index of the extent to which he has achieved the educational objectives set for him. Some questions are described as "catch questions." By this, it is usually meant that such questions suggest the wrong answer or involve an obscure restriction or qualification. In some cases the question calls for some unimportant detail or is ambiguous. There appears to have been no scientific investigation of the character of the examination questions asked of pupils. However, it is doubtless true that this criticism has justification in some cases because frequently teachers give relatively little time to the preparation of their questions, and these often reflect any hobbies or prejudices which the teachers may have. Experience in the construction of standardized educational tests has shown that it is difficult to eliminate all ambiguity and indefiniteness in questions. Hence, it is likely true that many questions are not well stated, and for this reason are not properly understood by those taking the examination. When this is the case, the "grades" tend to be inaccurate measures of achievement.

When an examination is set by some person other than the teacher of the class, it not infrequently happens that many of the questions pertain to topics which have received little or no attention during the instruction periods. In many schools it seems to be the custom for the superintendent or the principal, without consultation with the teacher in charge, to make out the questions for the final examination on which the pupils' semester grades are largely based. For

example, in a fifth-grade geography class which was brought to the attention of the writer, four of the five questions of the examination concerned current conditions about which the children, instructed only in their texts, knew little. A few pupils, fortunate enough to have heard these matters discussed in their own homes, received a passing grade. The majority of the class failed. This examination, interesting and in itself not subject to criticism, should not have been used, however, as a means for measuring the achievements of that particular class. It was not in agreement with the educational objectives toward which the teacher had directed their efforts. Such examinations are "hard" in the sense that capable students will answer only a relatively small per cent of the questions correctly, and are rightly criticized as being unjust because the students are not given an opportunity to demonstrate their achievements.

Two remedies for the imperfections of written examinations. There are other criticisms which might be enumerated, but our purpose here is only to show that the ordinary written examination is an imperfect measuring instrument. There are two ways of securing more accurate measures of the achievements of school children. We may make use of standardized educational tests and we may improve the examinations prepared by teachers. As the title implies, most of this book will be devoted to a consideration of certain standardized educational tests, but in the last chapter we shall return to the topic of written examinations and how they may be improved. This topic is, however, more important than the space devoted to it may indicate, because written examinations prepared by teachers will probably always be extensively used for securing measures of the achievements of school children. It is therefore important that teachers learn how to construct and administer them so that the measures they yield will be as accurate and useful as possible.

Standardized tests improved measuring instruments. Largely because we became conscious of the limitations of written examinations as instruments for measuring the abilities of school children, we have recently devised improved measuring instruments which are known as "standardized tests." In constructing standardized tests, the authors have attempted to eliminate or reduce to a minimum the defects of written examinations. In many standardized tests the pupil gives his answer by drawing a line under a word or phrase. In others only one answer can be accepted as correct. Detailed rules have been formulated for giving the standardized tests to the pupils and for marking the test papers. As a result standardized tests are more objective than ordinary written examinations. The measures of achievement depend very slightly, if at all, upon the person who administers a standardized test. Another teacher would have secured approximately the same results.

We describe these new measuring instruments as "standardized." By this we mean that objective norms have been ascertained so that we know the average scores which pupils belonging to a given grade or age group should make on each test. Thus there is uniformity in the administration of the test and also in the basis of interpreting the measures which it yields.

The "questions" for these standardized tests have been carefully formulated and selected. Usually they have been given to several hundred children in the process of the construction of the test. Those exercises which were found unsatisfactory for any reason have been discarded. In most cases the exercises have also been carefully selected with reference to the topics which they cover.

The structure of standardized tests is such that much time is saved for both teacher and pupils. The teacher has no questions to formulate or to write on the blackboard. A

minimum of writing is required of the pupils. In many cases they can answer the exercise by drawing a line or making a check mark. Much time is also saved in the marking of the papers. It is true that some standardized tests require considerable time, but in such cases the number of "questions" is large.

Standardized tests not perfect measuring instruments. Although standardized tests are distinctly superior in certain respects to ordinary written examinations, one should not think of them as being perfect measuring instruments. Some approach perfection more nearly than others, but even the best are subject to limitations which must not be overlooked. We shall point out the most significant of these defects in our discussion of the tests for particular subjects. However, we may note here (1) that the measures yielded by standardized tests are not absolutely accurate, in fact they frequently involve errors, both constant and variable, which are surprisingly large; (2) that in many cases conclusive evidence is lacking to show that the test measures what its title or a more explicit statement of its function claims it measures; (3) that we are as yet unable to measure directly many very important outcomes of instruction.

An intelligent attitude toward standardized tests. It is not intelligent to accept standardized tests as perfect measuring instruments, or even as approximating perfection. Not infrequently they are described as "scientific." The use of this descriptive term probably conveys to many persons a meaning which is not justified. Hence it should not be used unless it is defined or appropriately qualified. The measures yielded by standardized tests are frequently described as "reliable" or "accurate." These measures are in general more reliable than similar ones secured by ordinary examinations or by teachers' estimates, but they are far from being as reliable as our ordinary measures of length, area, volume,

14 EDUCATIONAL TESTS AND MEASUREMENTS

or weight. Compared with these they are distinctly unreliable. On the other hand, it is not intelligent to take the position that we cannot measure achievement and that standardized tests are no good. The intelligent attitude is between these two extremes. Standardized tests are valuable tools which teachers and supervisors may use in making their work more effective.

There seems to be a prevailing belief that, if certain methods of test construction are applied with sufficient elaborateness, we may expect to obtain a measuring instrument which will yield "scientifically accurate measures" of the specified abilities. Experience with our best tests shows that such a belief is not justified. There is not just *one* method of test construction. There are several, and the recognized principles of test construction applicable to the type of test one is constructing must be followed if his endeavors are to be successful.¹ Some methods involve only simple procedures. Others include complex statistical procedures, but the complexity of the procedure is no index of the value of the test.

The plan of the following chapters. In the following chapters the author has endeavored to keep in mind the needs of the teacher or student who is just beginning his acquaintance with standardized educational tests. Chapters II to VIII are devoted to a consideration of certain standardized tests for measuring achievements in different school subjects. General intelligence tests are described in Chapter IX. In these chapters there has been no attempt to include mention of all tests. In general a test has not been included unless it was known to be available for use; that is, could be purchased in quantity. This has eliminated certain tests whose publication has been discontinued and a much larger number which never advanced beyond the ex-

¹ See Chapter XI. Also Monroe, Walter S., *An Introduction to Educational Measurements*, chaps. II and IV. (Houghton Mifflin Company, 1922.)

perimental stage. In addition, a few new tests which are now in the experimental stage have been mentioned because they appear to be promising. A number of tests which the author does not value highly have been described because he realized that his judgment is doubtless influenced by the types of tests with which he has had most experience. The position has been taken that any test now available should be included among those described if it has received favorable recognition unless it appears reasonably certain that this recognition was not justified. The titles of the tests described in Chapters II to IX are summarized in Appendix B. The name of the publisher and also the price per hundred copies are included in this summary.

There has been no attempt to describe the construction of the various tests. A brief account of the general procedure of test construction, together with certain illustrations, is given in Chapter XI. In describing a test an effort has been made to cover the following points: (1) exact title, author, and grades in which the test is to be used; (2) number and character of exercises; ¹ (3) function of the test; (4) an evaluation with respect to objectivity, reliability, and validity; (5) norms. In the case of tests which appeared to be of minor importance and a number of newer tests for which relatively little information was available, this plan of description has been abbreviated. For a few of the more important tests certain phases of the description have been elaborated. In addition, the problem of measurement for each of the subjects has been emphasized and considerable space has been given to the general procedure of interpreting test scores and the remedial instruction which is appropriate for certain typical situations.

No final evaluation of the various tests has been attempted. One reason is that tests differ in function. The one which is

¹ This applies only to the more important and typical tests.

most suitable for a given purpose may not be the best one to use when considered with respect to another purpose. The author has attempted to give sufficient information about the tests described so that one may make an intelligent selection.

In Chapter X certain batteries of tests which have been grouped together are described under the title of "Testing Programs." Attention is also given to the questions arising in the planning of testing programs. Chapter XII is devoted to the "meaning of scores." A number of general questions arising in connection with the interpretation of test scores are considered here. In Chapters II to VIII the use of the information yielded by standardized tests is considered from the standpoint of instruction. Chapter XIII is devoted to the administrative and supervisory uses of educational tests. Chapter XIV presents suggestions for the improvement of measurements made by means of written examinations.

An attempt is made to define the various technical terms which are used in the following chapters when they are first presented. The more important of these technical terms are also defined in Appendix A. The arrangement is alphabetical and students should use this appendix as a reference when unfamiliar technical terms are encountered.

This book has been written with three general purposes in mind. The first of these is to assist the student in becoming acquainted with representative standardized educational tests. The second purpose, which is more fundamental, is to acquaint the student with the processes and technique of measurement so that he may improve the measurements which he makes by means of written examinations. In order to realize this purpose the problem of measurement has been emphasized and attention has been given to those factors of test construction which are suggestive of improve-

ment of written examinations. The third purpose relates to the use of educational measurements made either by means of standardized tests or by means of written examinations prepared by the teacher. Two uses have been recognized: (1) instructional, and (2) administrative and supervisory. The former will be of greater interest to teachers and the latter to principals and superintendents.

How to study the following chapters. In the study of Chapters II to IX, it is advisable for the student to concentrate his efforts upon a small number of representative tests. One who is relatively unacquainted with standardized tests will find that even the limited number of tests described appear as a bewildering array. An intensive study of a few typical tests is much more valuable than a superficial encyclopædic acquaintance with a large number of tests. It is desirable but not necessary to have at hand sample sets of tests which are to be studied intensively. These can be secured from the publishers given in Appendix B. When ordering sample sets one should specify that all directions and other accessories are desired. Although the cost of any one sample set is not large, the total for any considerable number will, in general, make it inadvisable to require students to provide themselves with sample sets of all the tests. When possible, some actual practice in the administration of one or two tests will be helpful. When no regular classes are available for this purpose, one or two tests may be given to the class in educational measurements.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What three methods may be employed for measuring abilities of school children?
2. Summarize the criticisms of written examinations. Which ones are based upon scientific evidence? Do any of these criticisms appear to be lacking in validity? If so, why?

18 EDUCATIONAL TESTS AND MEASUREMENTS

3. Distinguish between constant errors and variable errors.
4. What is the meaning of saying that the marking of examination papers is subjective?
5. What are standardized tests?
6. What attitude should one maintain toward standardized tests? Answer in detail.
7. Give evidence from your own experience to show that subjective norms are used in assigning grades to examination papers.

CHAPTER II

ARITHMETIC

I. THE PROBLEM OF MEASURING ARITHMETICAL ABILITIES

The problem of measuring a physical object. In measuring a physical object, such as a room, chair, haystack, or irregular-shaped field, it is necessary first to define one's purposes; that is, determine what dimensions are to be measured. For example, if our purpose is to ascertain the number of yards of carpet needed to cover a rectangular floor, only certain characteristics of the room, namely, length and width, are significant. On the other hand, if our purpose is to obtain a numerical measure of the lighting of the room, other characteristics, such as the number, position, and area of the windows, are essential. It is also necessary to secure an appropriate measuring instrument.

Our purpose in measuring arithmetical abilities. In general terms our purpose is to measure the results of our teaching; that is, the abilities which we are attempting to engender in our pupils. Another way of stating our purpose is to say that we wish to determine the extent to which our pupils have achieved the objectives set for them in the field of arithmetic. Numerous objectives have been recognized in the statements of the aim of teaching arithmetic, but it is generally agreed that among the desired outcomes of arithmetical instruction are the abilities required to perform the operations of addition, subtraction, multiplication, and division with integers and with fractions, both common and decimal. In addition there are certain objectives to be attained in connection with the solving of problems.

Arithmetical abilities automatic or habits. In order to be

properly equipped for the future work of the school as well as for the activities of an adult life, the pupil must be able to perform the operations of arithmetic rapidly and with a minimum of attention. As soon as he recognizes that a multiplication combination is called for, as, for example, 8×7 , the response, 56, must be forthcoming immediately. Time cannot be taken to think out the product. The pupil's attention must be reserved for deciding what operations to perform in dealing with the problems of arithmetic. The situation is similar to that which we have in any field of action where particular acts occur frequently and are always the same. Such acts must be reduced to the plane of habit if a person becomes skillful. We may therefore describe these arithmetical abilities as habits. Their functioning should be automatic.

Arithmetical abilities specific. A few years ago Stone¹ investigated the nature of ability in arithmetic and concluded that it is made up of a number of abilities which are relatively independent in their functioning. His conclusions have been corroborated by a number of other investigations² and it is now reasonably certain that, in teaching the operations of arithmetic, we are attempting to engender a number of specific abilities which are relatively distinct, and not a single arithmetical ability. It is obvious that a pupil might become expert in addition of integers, but not be able to do examples in long division. It also appears that the ability to add a column of three figures is not the same as the ability to add a column of twelve figures. In adding a column of figures it is necessary that one hold in mind the partial sum until he has added the next figure. This proc-

¹ Stone, C. W., *Arithmetical Abilities and Some Factors Determining Them*. (Teachers College Contributions to Education, no. 19, 1908.)

² Kallom, A. W., *Determining the Achievement of Pupils in Addition of Fractions*. (School Document no. 3, 1916. Boston Public Schools.)

ess must be repeated continuously until the final sum is reached, and a failure to do this will result in stopping the adding, at least temporarily. It is a frequent occurrence, for one who is not accustomed to adding long columns of figures, to find that he has stopped, perhaps has even lost the partial sum, and must begin again. The span of attention required in adding three figures is short, and pupils who are able to do examples of this type with a high degree of skill frequently are unable to add long columns of figures with an equal degree of skill. In fact, we have no reason to expect them to be able to do this type of example until they have practiced upon it. It has been said that there are as many different abilities as there are types of examples.

Separate types of examples in handling integers. Courtis,¹ the author of the Standard Research Tests in Arithmetic, has identified the following types of examples in the operations with integers:

Addition: (1) addition combinations; (2) single-column addition of three figures each; (3) "bridging the tens," as $38 + 7$; (4) column addition, seven figures; (5) carrying; (6) column addition with increased attention span, thirteen figures to the column; (7) addition of numbers of different lengths.

Subtraction: (1) subtraction combinations; (2) subtraction of 9 or less from a number of two digits, without "borrowing"; (3) same as the second, but with "borrowing"; (4) subtraction of numbers of two or more digits involving borrowing.

Multiplication: (1) multiplication combinations; (2) multiplicand two digits, multiplier one digit, and no carrying; (3) same as number 2, but with carrying; (4) long multiplication, without carrying; (5-8) zero difficulties, four types:

¹ Courtis, S. A., *Teacher's Manual for Courtis Standard Practice Tests*, (1916.)

22 EDUCATIONAL TESTS AND MEASUREMENTS

$$\begin{array}{r} 560 \\ \underline{47} \end{array}$$

$$\begin{array}{r} 807 \\ \underline{59} \end{array}$$

$$\begin{array}{r} 617 \\ \underline{508} \end{array}$$

$$\begin{array}{r} 753 \\ \underline{60} \end{array}$$

(9) long multiplication, with carrying.

Division: (1) division combinations; (2) simple division, no carrying; (3) same as number 2, but with carrying; (4) long division, no carrying; (5-6) zero difficulties, two cases:

$$\begin{array}{r} 690 \\ 71 \overline{)48990} \end{array}$$

$$\begin{array}{r} 302 \\ 31 \overline{)9362} \end{array}$$

(7) long division, with carrying, "first case, the first figure of the divisor is the trial divisor and the trial quotient is the true quotient":

$$\begin{array}{r} 72 \\ 63 \overline{)4536} \end{array}$$

(8) "second case, where the trial divisor is one larger than the first figure of the divisor, but the trial quotient is the true quotient":

$$\begin{array}{r} 63 \\ 49 \overline{)3087} \end{array}$$

(9) "third case, where the first figure of the divisor is the trial divisor, but the true quotient is one smaller than the trial quotient":

$$\begin{array}{r} 89 \\ 63 \overline{)5607} \end{array}$$

(10) "fourth case, where the first figure of the divisor must be increased by one to obtain a trial divisor and the second trial quotient must be increased by one to get the true quotient":

$$\begin{array}{r} 79 \\ 36 \overline{)2844} \end{array}$$

Each of these types of examples requires a different set of specific habits. To be sure, certain elements, such as the

fundamental combinations, are common elements, but careful analysis will show that the ability to do examples of one type is different from that required to do another. Not only will a careful analysis reveal this fact, but it has been repeatedly demonstrated by carefully conducted investigations. In addition to the specific habits which are required for the four fundamental operations with integers, a number of other specific habits are required for the types of examples in the field of fractions both common and decimal. At present we have only partial analysis of the examples in these fields, and for that reason it is not possible to list the types of examples that are within the range of school work. It appears likely that they are numerous.

A complete and detailed measurement of arithmetical skills would require that a test be provided for each type of example, but fortunately certain combinations can be made. An example in addition consisting of three columns of nine figures each includes the addition combinations, simple column addition, and carrying. Thus, if a pupil responds satisfactorily to examples of this type, we know that he possesses the ability to do the types of addition examples involved therein. On the other hand, if his response to this type of example is unsatisfactory, we do not know just what elemental ability he lacks. The use of a single test of this type to measure a group of arithmetical abilities has this very obvious limitation in diagnosing the conditions which exist, but it does provide a very satisfactory general survey.

Significant characteristics of arithmetical abilities in the fundamental operation. We recognize two types of objectives in teaching the fundamental operations of arithmetic. Pupils are expected to increase in skill in doing certain types of examples. For example, when a pupil first learns to do long columns in addition with carrying, he will

24 EDUCATIONAL TESTS AND MEASUREMENTS

work slowly and is likely to make a relatively large number of errors. As he is given practice on this type of example, he increases both his rate of work and its accuracy. We say that he is growing in "skill" or "fluency." However, the pupil is at the same time developing in another way. He is learning to do other types of examples which are more difficult. We may describe this growth by saying that he is increasing in "power" to do difficult examples. Thus we have two objectives, "skill" and "power."

The significant characteristics of skill are "rate of work" and "accuracy." Thus, if we wish to measure the skill of pupils in doing certain types of examples, it is necessary to use a test which yields measures of rate (number of examples attempted in a given time) and accuracy (per cent of examples done correctly). These two characteristics or dimensions are taken as a measure of a pupil's ability to do a given type of example.¹ Such measuring instruments are commonly called "rate tests." On the other hand, if we wish to secure a measure of pupils' "power," we should use a test (scale) which will tell us how difficult examples each pupil is just barely able to do correctly. In this case the rate of work is not considered. Measuring instruments of this type are called "power tests." It should be noted that our present tendency is to eliminate the more difficult and intricate examples from the curriculum. Hence there are certain limitations placed upon the objective of "power" in arithmetic.

¹ Strictly speaking, the number of examples done and the per cent of examples correct is a measure of the pupil's performance rather than of his ability. A pupil's performance is affected by many factors such as his emotional status, physical condition, light, temperature, and the like. Or, it may be that a pupil does not try to do his best on a given test. A pupil's ability can only be inferred from his performance, but when conditions are properly controlled, such inference is reliable in all except a few cases. In order to avoid an awkward form of statement and because the practice is general, we shall speak of a score as a measure of a pupil's ability.

Significant characteristics of the ability to solve problems. Many elements enter into the ability to solve problems.¹ In the case of written problems, the statement must be read in defining the problem. Then facts and principles must be recalled as a basis for formulating a plan of solution. This plan of solution may be verified by comparison with the statement of the problem and other facts and principles. The calculations performed in executing the "plan of solution" for a problem is not a part of the reasoning process. Thus a pupil may reason correctly and fail to obtain the correct answer to a problem because he copied some numbers incorrectly or made an error in his calculations. Hence, if we mean by the "ability to solve problems" the ability to decide upon the correct plan of solution, it follows that the correctness of the answer is not an important characteristic of this ability. The significant characteristics describe the pupil's ability to formulate a correct plan of solution. Other things being equal, the pupil who can most quickly decide upon the plans of solution for a list of problems possesses the highest degree of ability in this field. However, the rate of work appears to be relatively much less important than the accuracy of the plans. Most reasoning tests do not yield measures of rate of work.

It appears likely that types of problems exist, but their distinguishing characteristics are not as obvious as in the case of the types of examples described on page 21. A pupil grows in ability to solve problems of a given type and also in ability to solve more complex and difficult types. The latter type of growth which we may call "power to solve

¹ The word "problem" is used by some writers to designate both "examples" and "problems." In this book the word "example" will be used to designate exercises which explicitly call for certain arithmetical operations. The word "problem" will designate only those exercises which require the pupil to determine first what operations are to be performed.

problems" appears to be more important than increase in "power" in the operations of arithmetic. It is, however, necessary to bear in mind that problems are not included in our educational objectives on the basis of their difficulty. They are included because of their social importance and their difficulty is merely incidental.

General tests vs. diagnostic tests. In considering the tests described in the following pages it is necessary to keep in mind that all do not have the same function. Some have a general function, others are diagnostic. A general survey test furnishes general information. Such information is useful in determining the general effectiveness of the instruction. Hence the superintendent or principal will frequently find a general test best suited to his purposes. The teacher, however, is primarily concerned with details of instruction and with individual pupils, and therefore must have detailed information in order to know how to adjust the instruction to the needs of the individual pupils. She needs to learn what types of examples her pupils can do with a satisfactory degree of fluency, and what types they cannot do. She needs to learn what pupils possess standard ability and what pupils do not. A general test serves to locate the pupils who are not yet up to standard, but a more elaborate test must be used to reveal the exact nature of the shortcomings of the pupils. Hence teachers will find the diagnostic tests most helpful.

II. STANDARDIZED TESTS FOR MEASURING ABILITIES IN THE OPERATIONS OF ARITHMETIC

General structure of tests upon the operations of arithmetic. The different tests upon the operations of arithmetic exhibit many common characteristics of general structure. They consist of examples which are usually printed so that the pupil can do them without copying the numbers.

He is asked only to do the examples.¹ The tests differ in the types of the examples used and in their arrangement. In most cases there are two or more sub-tests in a series. In some series a given sub-test is limited to a single type of example. In others there is a wide range of types within a sub-test which may be devoted to a single operation.

1. The Courtis Standard Research Tests, Series B²

General structure. The Standard Research Tests, Series B, or as they are commonly called, the Courtis Arithmetic Tests, were constructed by S. A. Courtis, of the Detroit Public Schools, during the year of 1913-14.³ They have probably been more widely used than any other instrument for measuring arithmetical abilities, and as a result we have norms based upon a very large number of scores. The series consists of four tests, printed as a four-page folder, each test occupying a page. The tests are accompanied by detailed directions which not only make their use easier, but also tend to insure that different persons will administer them in the same way.

Test No. 1. Addition

The twenty-four examples of this test have been constructed so that all have the same form, three columns of nine figures each. The following are samples of the examples. Time allowed, 8 minutes.

¹ In the case of examples which require little written calculation there has been some experimentation with exercises in which the pupil is given several answers to the example and is asked to choose the correct one.

² This series of tests will be described in considerable detail, partly because it is important, but primarily because it is the first series of tests to be considered. Many of the concepts presented here will be helpful to the reader in understanding the briefer discussion of other tests.

³ Earlier, Courtis devised another set of arithmetic tests which were known as Series A. The publication of this first series has been discontinued.

28 EDUCATIONAL TESTS AND MEASUREMENTS

927	297	136	486	384	176
379	925	340	765	477	783
756	473	988	524	881	697
837	983	386	140	266	200
924	315	353	812	679	366
110	661	904	466	241	851
854	794	547	355	796	535
965	177	192	834	850	323
<u>344</u>	<u>124</u>	<u>439</u>	<u>567</u>	<u>733</u>	<u>229</u>

In giving the test the pupils are directed as follows:

You will be given eight minutes to find the answers to as many of these addition examples as possible. Write the answers on this paper directly underneath the examples. You are not expected to be able to do them all. You will be marked for both speed and accuracy, but it is more important to have your answers right than to try a great many examples.

The directions for the other three tests are similar.

Test No. 2. Subtraction

This test consists of twenty-four examples, each involving the same number of subtractions. The following are samples. Time allowed, 4 minutes.

107795491	75088824	91500053	87939983
<u>77197029</u>	<u>57406394</u>	<u>19901563</u>	<u>72207316</u>

Test No. 3. Multiplication

This test consists of twenty-four examples of this type. Time allowed 6 minutes.

8246	3597	5739	2648	9537
<u>29</u>	<u>73</u>	<u>85</u>	<u>46</u>	<u>92</u>

Test No. 4. Division

This test consists of twenty-four examples of this type. Time allowed, 8 minutes.

25)6775	94)85352	37)9990	86)80066
<u>73)58765</u>	<u>49)31409</u>	<u>68)43520</u>	<u>52)44252</u>

Marking the papers. In marking the test papers, which is done by the use of a printed answer card which is placed just below the pupil's answers, no credit is given for examples partly right nor for examples partly completed. A pupil's score is the number of examples attempted and the number right. This simple plan of marking the papers insures uniformity. In contrast with the marking of the examination papers described on page 5, the scoring of these tests is highly objective.

Each of the examples of a test calls for the same number of operations under approximately the same conditions. This makes the examples of each test approximately equal in difficulty. Any example of the addition test, say the seventh, is just as difficult as any other, say the second. Thus, the tests consist of twenty-four equal units, just as a yardstick consists of thirty-six units (inches). The measure of a pupil's ability is represented by the distance he advances along the scale in the given time; that is, by the number of examples done and by the per cent of these examples which have been done correctly.

Since an example of one of these tests is defined as so many operations under certain conditions, it is possible to construct other tests equal in difficulty. Four duplicate forms have been constructed. This makes it possible to use a different form when the tests are repeated.

Recording the scores of a class. A portion of the class record sheet for recording the scores of a class is shown in Fig. 2. This figure contains merely the blank for addition, but those for the other three tests of the series are identical with it. Detailed instructions for recording scores are printed on the record sheet. The large figures at the top of Fig. 2 refer to the number of examples attempted and the small figures within the squares refer to the number of examples done correctly. The sheet is arranged so that the per cent of

examples done correctly is computed automatically, and the distribution of the scores according to both rate and accuracy is obtained at the same time. The scores of a seventh-grade class are shown in Fig. 2. The numbers written in certain of the squares represent the number of pupils whose scores fell within these divisions of the record sheet. The distribution according to rate is found at the bottom of the record sheet and is to be read thus: Three pupils attempted only six examples, two pupils attempted only seven examples, five pupils attempted only eight examples, etc. The distribution according to accuracy is found at the right-hand side of the sheet and is to be read thus: The per cent of examples done correctly by two pupils was less than fifty per cent, for five pupils it was between sixty per cent and seventy per cent, etc. These two distributions, the one according to the per cent of examples done correctly and the other according to the number of the examples attempted, describe the ability of the pupils of this class to do the examples of the addition test.

Function. The Courtis Standard Research Tests, Series B, are diagnostic to the extent of yielding separate measures for each of the operations with integers, but they are general tests within the field of each operation. They measure skill rather than power.

Norms. The degree of ability which typical pupils belonging to a given school grade should possess is called a norm for that grade. Norms are necessary to give meaning to the scores which pupils make. In most cases the norms are median¹ or average scores and thus represent merely the consensus of present practice. Such norms are open to the criticism that we cannot be certain that our present practice is satisfactory, but it seems probable that norms derived in this way will not be changed materially in the near future

¹ See Appendix A for definition of "median."

Test No. 1. Addition. Standard Scores: Grade 3, 4; Gd 4, 6; Gd 5, 8; Gd 6, 10; Gd 7, 11; Gd 8, 12. Accuracy 100%

Score		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total	%
100		—	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	6	
90				—	—	—	—	—	—	—	—	2	2	11	12	13	14	15	16	17	18	19	20	21	22	23	4	
80				—	—	—	4	6	6	7	8	4	3	10	11	12	13	14	15	16	17	18	19	20	21	22	12	
70				—	—	3	—	—	5	6	7	7	1	9	10	11	12	13	14	15	16	17	18	19	20	21	5	
60				—	2	—	3	4	—	6	6	6	7	8	8-9	9-10	10-11	11	12	13	14	15	16	17	18	19	5	
50			1	—	—	2	—	3	4	4	5	6	6	7	8	9	10	11	12	13	14	15	16	17	18	19	5	
0 to 49		0	0	0	0-1	0-1	0-2	0-2	0-3	0-3	0-4	0-4	0-5	0-6	0-6	0-6	0-7	0-7	0-8	0-8	0-9	0-9	0-10	0-10	0-11	0-11	2	
Total								3	2	5	4	8	6	3	5	1	1	1	1								39	
Median Scores:																												
Rate																												
Accuracy																												
Efficiency																												
Standard Scores																												

For instructions, see other side of this page.

FIG. 2. SHOWING FORM OF TABULATION SHEET FOR RECORDING SCORES OBTAINED BY USING THE COURTIS STANDARD RESEARCH TESTS, SERIES B, AND THE SCORES OF A SEVENTH-GRADE CLASS IN ADDITION

provided they are based upon a sufficient number of typical cases. The topic of norms and the use of them is discussed at length in Chapter XII, and the reader may profitably study it in connection with the norms presented in this and the following chapters. The reader should also read the discussion of the "accuracy of individual scores" on pages 69-72.

In Table I there are given three sets of grade norms: (1) general median scores based upon distributions of "many thousands of individual scores in tests given in May or June, 1915-16. The distribution for each grade was made up of approximately equal numbers of classes from large-city schools and from small-city and country schools"; (2) the norms proposed by Courtis after three years' use of these tests; (3) Boston median scores after the tests had been used for three years.

With reference to the norms which he has proposed Courtis says:

The speeds set as standard are approximately the average speeds at which the children of the different grades have been found to work when tested at the end of the year, when for any one grade a random selection of five thousand scores from children in schools of all types and kinds are used as a basis of judgment.

Standard accuracy is perfect work, one hundred per cent. This is a tentative standard only, as there is available very little information in regard to the factors that determine accuracy and the effects of more efficient training.

At present in addition and multiplication it is only very exceptional work in which the median rises above eighty per cent accuracy, while in subtraction and division the limiting level is ninety per cent.

Standard speeds are not likely to change greatly. Standard accuracy is surely destined to approach much more nearly one hundred per cent than present work would indicate.

Standard scores are not only goals to be reached; they are limits not to be exceeded. It seems as foolish to overtrain a child

TABLE I. GRADE NORMS FOR COURTIS STANDARD RESEARCH TESTS, SERIES B

GRADE		ADDITION		SUBTRACTION		MULTIPLICATION		DIVISION	
		Rate	Accuracy	Rate	Accuracy	Rate	Accuracy	Rate	Accuracy
IV..	General	7.4	64	7.4	80	6.2	67	4.6	57
	Courtis	6	100	7	100	6	100	4	100
	Boston	8	70	7	80	6	60	4	60
V..	General	8.6	70	9.0	83	7.5	75	6.1	77
	Courtis	8	100	9	100	8	100	6	100
	Boston	9	70	9	80	7	70	6	70
VI..	General	9.8	73	10.3	85	9.1	78	8.2	87
	Courtis	10	100	11	100	9	100	8	100
	Boston	10	70	10	90	9	80	8	80
VII..	General	10.9	75	11.6	86	10.2	80	9.6	90
	Courtis	11	100	12	100	10	100	10	100
	Boston	11	80	11	90	10	80	10	90
VIII..	General	11.6	76	12.9	87	11.5	81	10.7	91
	Courtis	12	100	13	100	11	100	11	100
	Boston	12	80	12	90	11	80	11	90

Rate is the number of examples done in the time allowed.

Accuracy is the per cent of examples correct.

"General" medians were determined by Courtis on the basis of the 1916 tabulations and summaries of tabulations of other years. Courtis, S. A., *Third, Fourth, and Fifth Annual Accountings, 1913-16*. (Department of Coöperative Research, Detroit.)

The Boston norms were established after using the tests for three years. Ballou, F. W., *Arithmetic, the Courtis Standard Tests in Boston, 1912-15*. (Bulletin 10 of the Department of Educational Investigation and Measurement.)

as it is to undertrain him. All direct drill work should, in the judgment of the writer, be discontinued once the individual has reached standard levels. If his abilities develop further through incidental training, well and good, but the superintendent who, by repeated raising of standards, forces teachers and pupils to spend each year a larger percentage of time and effort upon the mere

mechanical skills, makes as serious a mistake as the superintendent who is too lax in his standards.¹

Comparisons with these norms or any others are valid only when the tests have been given under standard conditions. Slight changes in the method of giving the tests may affect the scores as much as the difference in the norms from one grade to another.

2. *Courtis Standard Supervisory Tests in Arithmetic*

Purpose. As the name of these tests implies, they are intended to yield very general measures of arithmetical abilities and are to be used for supervisory rather than instructional purposes. Examples in each of the four operations with integers are included in both tests. Test A, which is recommended for Grades IV B to V B, consists of relatively simple examples. Those of Test B, designed for Grades V A to VIII A, are more difficult. The arrangement is such that a pupil completes the examples of a given operation before taking up the next. Since only the entire test is timed, those pupils who work slowly will not have an opportunity to try the examples in division, and in some cases those in multiplication will not be reached. The time allowance is varied from grade to grade so that pupils who have attained standard ability will be able to finish the test. There are several forms of each test, but no determination of its reliability² has been reported. Each test is printed on a simple sheet and hence is relatively inexpensive.

Each pupil is given two scores, one, "Number of samples tried," and the other, "Examples right." In assembling the scores of a class the possible combinations of these two scores are grouped under five heads: Group I, children of

¹ Courtis, S. A., *Third, Fourth, and Fifth Annual Accountings, 1913-16*, p. 49. (Department of Coöperative Research, Detroit.)

² See Appendix A for definition of "reliability."

standard ability; Group II, children for whom regular work will furnish sufficient drill; Group III, children in need of thorough drill; Group IV, children who need special attention and extra drill; and Group V, children for whom some special adjustment of work must be made. A class score is found by calculating the per cent of pupils falling in each group. These per cents are then multiplied by arbitrarily assigned numbers, 10, 9, 7, 4, and 0 respectively. The sum of these products is the class score. The maximum class score will occur when 100 per cent of the class are in Group I. In this case the class score would be 100×10 , or 1000. The city scores for Detroit for 1920-21 are given below.

GRADE	TEST A			TEST B						
	IV B	IV A	V B	V A	VI B	VI A	VII B	VII A	VIII B	VIII A
September, 1920	98	383	575	410	496	563	591	631	650	730
January, 1921 (before promotion)	635	806	845	759	808	828	825	831	852	891
February, 1921 (after promotion)	192	611	763	550	639	675	702	722	779	775
May, 1921	641	810	855	715	771	791	803	815	846	839

The plan of tabulating the scores of a class implies a general instructional function. However, the structure of test suggests that the grouping of the members of a class should not be considered highly accurate. Furthermore, it is obvious that a pupil's score is not an indication of the operation or types of examples on which he needs drill. Hence the instructional function of the Courtis Supervisory Tests in Arithmetic is very limited, although they will doubtless be found helpful for a preliminary general survey.

3. *The Cleveland-Survey Arithmetic Tests*

General structure. This series of tests originated in the survey of the Cleveland Public Schools and is designed to be used in Grades III B to VIII A, although pupils in the lower grades will not be able to do the more difficult examples. Contrary to the implication of the title, the function is diagnostic. There are fifteen sub-tests, each one being devoted to one type or at most to a limited number of types of examples. Samples of these sub-tests are given on pages 37-39. Each test is timed separately, the allowances being as follows:

Set A.....30 seconds	Set F..... 1 minute	Set K.....2 minutes
Set B.....30 seconds	Set G..... 1 minute	Set L.....3 minutes
Set C.....30 seconds	Set H.....30 seconds	Set M.....3 minutes
Set D.....30 seconds	Set I..... 1 minute	Set N.....3 minutes
Set E.....30 seconds	Set J..... 2 minutes	Set O.....3 minutes

As in the case of the Curtis Standard Research Tests, Series B, the examples of each test are approximately equal in difficulty. Thus each test may be considered to consist of approximately equal units. In marking the test papers, no credit is given for examples partly right nor for examples partly completed which insures uniformity in scoring. A pupil's score is the number of examples attempted and the number right.

Norms. These tests have been widely used, but there has been no elaborate compilation of scores for the purpose of arriving at norms. The median scores for St. Louis, Missouri, and Grand Rapids, Michigan, are given in Table II. These will be useful for comparative purposes, but it should be noted that since the function is primarily diagnostic precise norms are not as essential as in the case of tests which have a general or survey function.

Limitations as diagnostic tests. In considering the com-

pleteness of this series of tests, it must be remembered that decimal fractions are omitted, and that two tests are certainly inadequate for the field of common fractions. Furthermore, all of the types of examples with integers listed on pages 21-22, are not represented. These tests, however, furnish a means for securing more detailed measurements of the arithmetical abilities of pupils than are possible by using the Courtis Standard Research Tests, Series B.

It should be noted also that coaching, or even recent instruction upon particular types of examples, is likely to increase materially the scores upon the tests of this series. For example, pupils who have had recent instruction on common fractions will make relatively higher scores on the fraction tests than they will make upon the rest of the series and also higher than they may make a little later in the year. Since the time allowances are short, slight errors in timing will affect the scores. It is necessary to bear these limitations in mind when using the Cleveland-Survey Tests.

Set A. Addition

$$\begin{array}{r} 1 \quad 6 \quad 9 \quad 0 \quad 4 \quad 1 \quad 7 \quad 9 \quad 3 \quad 2 \quad 1 \quad 3 \quad 6 \\ 2 \quad 6 \quad 5 \quad 1 \quad 2 \quad 3 \quad 7 \quad 6 \quad 0 \quad 4 \quad 5 \quad 8 \quad 9 \\ \hline \end{array}$$

Set B. Subtraction

$$\begin{array}{r} 9 \quad 7 \quad 11 \quad 8 \quad 12 \quad 1 \quad 9 \quad 13 \quad 4 \quad 12 \\ 9 \quad 3 \quad 6 \quad 1 \quad 3 \quad 0 \quad 7 \quad 8 \quad 3 \quad 6 \\ \hline \end{array}$$

Set C. Multiplication

$$\begin{array}{r} 2 \quad 4 \quad 9 \quad 0 \quad 5 \quad 4 \quad 2 \quad 7 \quad 4 \quad 9 \\ 2 \quad 7 \quad 8 \quad 2 \quad 6 \quad 1 \quad 9 \quad 6 \quad 0 \quad 5 \\ \hline \end{array}$$

Set D. Division

$$\begin{array}{r} 3 \overline{)9} \quad 4 \overline{)32} \quad 6 \overline{)36} \quad 2 \overline{)0} \quad 7 \overline{)28} \quad 9 \overline{)9} \quad 3 \overline{)21} \end{array}$$

Set E. Addition

5	2	9	2	6	1	4	9
2	8	8	8	3	4	6	7
2	8	0	5	4	2	5	1
0	5	7	0	8	5	3	5
<u>4</u>	<u>1</u>	<u>6</u>	<u>6</u>	<u>8</u>	<u>4</u>	<u>4</u>	<u>3</u>

Set F. Subtraction

616	1248	1365	1092	716
<u>456</u>	<u>709</u>	<u>618</u>	<u>472</u>	<u>344</u>

Set G. Multiplication

2345	9735	8642	6789	2345
<u>2</u>	<u>5</u>	<u>9</u>	<u>2</u>	<u>6</u>

Set H. Fractions

$$\frac{3}{5} + \frac{1}{5} =$$

$$\frac{6}{9} - \frac{4}{9} =$$

$$\frac{4}{9} + \frac{1}{9} =$$

$$\frac{8}{9} - \frac{7}{9} =$$

Set I. Division

$$4 \overline{)55424}$$

$$7 \overline{)65982}$$

$$2 \overline{)58748}$$

$$5 \overline{)41780}$$

Set J. Addition

7	9	4	7	2	9	6	7	7	8	9	4	3	2
5	2	5	1	9	6	9	1	8	0	5	3	1	1
4	4	8	9	4	2	6	5	5	7	3	7	7	6
2	8	1	4	8	4	7	1	4	1	4	7	6	6
6	2	4	3	5	7	0	4	1	8	6	0	9	1
0	7	8	2	1	1	4	6	8	5	2	2	6	8
5	5	5	8	5	3	3	5	2	1	3	9	3	6
1	3	1	5	2	9	7	3	1	3	9	5	4	9
8	6	3	2	4	2	1	3	3	7	2	6	5	7
3	1	9	7	3	3	6	7	9	4	2	3	4	5
2	4	6	7	6	8	0	6	8	9	8	4	2	2
9	8	3	1	7	5	6	1	4	4	5	8	9	2
<u>9</u>	<u>8</u>	<u>5</u>	<u>9</u>	<u>6</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>5</u>	<u>4</u>	<u>6</u>	<u>8</u>	<u>9</u>	<u>4</u>

Set K. Division

$$21 \overline{)441}$$

$$32 \overline{)672}$$

$$23 \overline{)483}$$

$$51 \overline{)1173}$$

Set L. Multiplication

$$\begin{array}{r} 8246 \\ \times 29 \\ \hline \end{array}$$

$$\begin{array}{r} 3597 \\ \times 73 \\ \hline \end{array}$$

$$\begin{array}{r} 5739 \\ \times 85 \\ \hline \end{array}$$

$$\begin{array}{r} 2648 \\ \times 46 \\ \hline \end{array}$$

$$\begin{array}{r} 9537 \\ \times 92 \\ \hline \end{array}$$

Set M. Addition

7493	8937	8625	2123	5142	3691
9016	6345	4091	1679	0376	4526
6487	2783	3844	5555	4955	7479
7591	4883	8697	6331	9314	2087
6166	1341	7314	6808	5507	8165

Set N. Division

$$67 \overline{)32763}$$

$$48 \overline{)28464}$$

$$97 \overline{)36084}$$

$$59 \overline{)29382}$$

Set O. Fractions

$$\frac{11}{15} + \frac{1}{6} =$$

$$\frac{9}{14} - \frac{1}{4} =$$

$$\frac{3}{4} \times \frac{5}{6} =$$

$$\frac{20}{21} \div \frac{1}{6} =$$

TABLE II. GRADE NORMS FOR THE CLEVELAND-SURVEY
ARITHMETIC TESTS — MEDIAN NUMBER OF EXAMPLES CORRECT
St. Louis, Missouri

TEST	GRADES											
	III B	III A	IV B	IV A	V B	V A	VI B	VI A	VII B	VII A	VIII B	VIII A
A	14.6	18.3	19.8	21.3	22.5	22.5	26.3	26.4	27.8	28.4	32.3	32.2
B	9.9	12.2	17.1	17.0	18.0	20.0	20.3	20.6	22.8	24.2	26.7	28.3
C	7.6	10.5	16.7	15.4	16.9	16.7	18.2	18.3	18.9	19.8	20.7	21.9
D	9.0	12.2	15.8	16.3	18.4	17.8	19.3	20.5	21.3	22.3	23.8	25.7
E	3.8	4.8	5.7	5.4	6.0	6.1	6.9	7.1	6.6	7.4	8.0	8.4
F	2.3	3.5	5.6	6.0	6.4	7.4	8.0	8.3	8.5	9.6	10.1	11.3
G	2.7	3.5	4.9	5.1	5.5	5.6	5.9	6.2	6.4	6.9	7.4	7.8
H	0.7	3.8	7.8	6.8	4.8	6.5	8.0	8.1	9.5	9.7	10.8	12.0
I	1.1	1.4	2.0	2.0	3.0	3.2	3.9	4.1	4.5	5.0	5.4	5.8
J	1.6	2.9	3.8	3.9	4.1	4.3	5.0	5.1	5.2	5.3	5.4	5.8
K			3.3	4.0	5.0	5.8	6.9	7.4	8.3	9.7	10.3	11.7
L			2.5	2.9	3.1	3.4	4.3	4.1	4.6	4.7	5.2	5.3
M	0.5	2.1	2.9	3.3	3.4	3.7	4.2	4.4	4.5	4.9	5.2	5.3
N			0.8	1.1	1.3	1.4	1.6	1.8	2.0	2.0	2.6	2.7
O				2.5	3.3	3.3	3.6	4.1	4.8	5.6	6.1	6.6

TABLE II.—*Continued*
Grand Rapids, Michigan

TEST	GRADES											
	III B	III A	IV B	IV A	V B	V A	VI B	VI A	VII B	VII A	VIII B	VIII A
A	11.8	13.4	13.6	16.4	20.3	21.5	22.8	25.0	26.5	27.3	29.5	30.3
B	6.3	8.4	9.1	12.1	14.7	15.9	16.8	19.1	21.3	20.7	22.8	25.5
C			7.1	11.3	13.7	14.0	15.5	17.0	17.7	18.8	19.3	20.7
D			6.9	10.4	12.5	14.3	15.5	16.9	18.4	19.7	20.5	23.0
E			4.1	4.6	5.2	5.4	6.0	6.6	7.2	7.2	7.8	8.1
F			2.8	4.1	6.0	6.5	7.1	8.0	9.3	9.6	10.3	11.0
G			2.2	3.3	4.5	4.9	5.3	5.6	6.1	6.1	6.7	6.8
H						6.3	6.2	6.5	9.0	7.8	8.6	8.8
I			0.7	0.9	1.3	1.4	2.3	3.0	3.8	4.1	4.0	4.7
J				2.8	3.4	3.7	4.1	4.5	5.4	5.3	5.7	6.5
K					3.0	4.3	5.4	6.5	7.5	8.8	9.7	10.3
L					2.3	2.9	3.3	3.6	4.3	4.5	4.9	4.9
M				2.3	3.0	3.6	4.3	4.5	4.9	5.0	5.7	5.7
N					0.7	0.8	1.1	1.4	1.7	1.8	2.0	2.3
O							3.5	3.6	3.9	4.6	5.5	4.8

4. *The Lunceford Diagnostic Tests in Addition*

Purpose. These tests are designed for use in the primary grades. Test I consists of fifty-four addition combinations in example form:

3	7	2	8	9	6	8	7	4
2	4	2	5	2	4	0	3	1
—	—	—	—	—	—	—	—	—

Test II consists of the same combinations, but the order of the numbers is inverted. The time limit is varied in the different grades. A feature of the tests is a "diagnostic record sheet." Each combination of the two tests appears on this sheet and a record for a class is made by recording the total number of errors for each combination. Such a diagnosis tells a teacher the exact combinations on which the class as a whole need instruction. If these combinations "have been taught," the remedial instruction will consist largely of drill. A similar diagnosis of an individual pupil can be obtained

from his test paper. Similar diagnostic tests could be constructed for each of the other operations.

It should be noted that these tests are not measuring instruments in the usual sense. It is true that each pupil can be given a score, the number of combinations given correctly, or even two scores as in the case of the Courtis Standard Research Tests, Series B. The scores for a class could be assembled in the usual way and the median, or average, calculated as a class score. This, however, would destroy the specific function of these tests. The author has not provided for a class score, and for this reason there are no norms other than those implied in the structure and function of the test. Pupils are expected to be able to give the correct answer to *all* of the combinations in the time allowed.

5. Monroe's Diagnostic Tests in Arithmetic

General structure. This is a series of twenty-one tests. It is similar in general structure to the Cleveland-Survey Tests in Arithmetic, but differs from them in certain respects. There are no tests on the fundamental combinations. There are five instead of two tests on common fractions and five tests upon decimal fractions. Only one form has been published, but other forms could be easily constructed. The tests are printed on four separate four-page folders. The tests printed together in a folder are called a "part." Tests I to XI (Parts I and II) are confined to integers and are designed for use in Grades IV to VIII. Part III, common fractions, is designed for Grades V to VIII. Tests XVII to XXI (Part IV) are limited to multiplication and division of decimals. They are to be used in Grades VI to VIII. Each test is timed separately, and as in the case of the Cleveland-Survey Tests most of the time allowances are short. Only thirty seconds are allowed for each of the tests upon decimal fractions, but it should be noted that the pupil is required

only to place the decimal point in the answer which is printed on the test. The total working time for the twenty-one tests is thirty-one minutes. A pupil's score is the number of examples done correctly.

The following samples will illustrate the types of examples included in the several tests:

ADDITION

<i>Test I</i>	<i>Test V</i>	<i>Test VII</i>	<i>Test XII</i>	<i>Test XV</i>
4	7862	7	$\frac{1}{6} + \frac{1}{3} =$	$\frac{1}{6} + \frac{3}{5} =$
7	5013	6	$\frac{5}{6} + \frac{1}{2} =$	$\frac{3}{12} + \frac{5}{8} =$
2	1761	6	$\frac{3}{10} + \frac{3}{5} =$	
-	5872	5	$\frac{5}{9} + \frac{2}{3} =$	
	<u>3739</u>	0		
		5		
		1		
		8		
		7		
		3		
		3		
		1		
		2		
		-		

SUBTRACTION

<i>Test II</i>	<i>Test IX</i>	<i>Test XIII</i>
37 94	739 1853	$\frac{3}{4} - \frac{2}{5}$
<u>5</u> <u>8</u>	<u>367</u> <u>948</u>	$\frac{5}{6} - \frac{3}{4}$

MULTIPLICATION

<i>Test III</i>	<i>Test VIII</i>	<i>Test X</i>
6572	4857	560 807 617 840
<u>6</u>	<u>36</u>	<u>37</u> <u>59</u> <u>508</u> <u>80</u>
<i>Test XIV</i>	<i>Test XVIII</i>	<i>Test XX</i>
$\frac{2}{3} \times \frac{3}{4}$	657.2 67.50	487.5 57.28
$\frac{2}{5} \times \frac{3}{7}$	<u>.7</u> <u>.03</u>	<u>.62</u> <u>9.5</u>
$\frac{5}{12} \times \frac{3}{5}$	46004 20250	302250 544160

In Tests XVIII and XX the pupil is simply to insert the decimal point in the product which is given. In the samples only the variations in the multiplier are given. Each multiplier is used with three types of multiplicands (657.2, 65.72, 6.572). Thus each test includes six types of examples.

DIVISION

Test IV

$$8 \overline{)3840}$$

Test VI

$$82 \overline{)3854}$$

Test XI

$$47 \overline{)27589}$$

Test XVI

$$\frac{2}{3} \div \frac{1}{3}$$

$$\frac{4}{7} \div \frac{2}{3}$$

$$\frac{3}{8} \div \frac{2}{3}$$

Test XIX

$$.4 \overline{)148} \text{ Ans.: } 37$$

$$.9 \overline{)65.7} \text{ Ans.: } 73$$

$$.6 \overline{)1.68} \text{ Ans.: } 28$$

$$.7 \overline{).301} \text{ Ans.: } 43$$

Test XVII

$$.03 \overline{)16.2} \text{ Ans.: } 54$$

$$.07 \overline{)1.82} \text{ Ans.: } 26$$

$$.05 \overline{).415} \text{ Ans.: } 83$$

$$.06 \overline{)7.44} \text{ Ans.: } 124$$

Test XXI

$$.47 \overline{)2758.9} \text{ Ans.: } 587$$

$$8.2 \overline{)38.54} \text{ Ans.: } 47$$

$$79 \overline{)36.893} \text{ Ans.: } 467$$

Test XI is a composite test involving the four "cases" of long division given by Courtis. In Tests XVII, XIX, and XXI the pupil is to write the answer in the proper place and insert the decimal point. In Test XXI each of the three types of divisors is placed with each of four types of dividends, thus providing twelve types of examples.

Function. As the title implies, these tests are intended to fulfill a diagnostic function. They yield twenty-one separate measures of arithmetical abilities. Most of the tests are confined to a single type of example. Skill rather than "power" is measured. By combining the scores yielded by the separate tests into a single score, a survey function would be fulfilled, but this is not recommended, since this function can be more easily realized by tests designed for this purpose.¹

Norms. As we pointed out in our discussion of the Cleveland-Survey Tests, precise norms are less necessary for diagnostic purposes than for supervisory or administrative uses.

¹ See Monroe's General Survey Scales in Arithmetic, pages 45-49.

44 EDUCATIONAL TESTS AND MEASUREMENTS

The norms given in Table III are only approximate, but probably will be found sufficient for diagnostic purposes. Such norms constitute a set of detailed objectives for the por-

TABLE III. MONROE'S DIAGNOSTIC TESTS IN ARITHMETIC —
GRADE MEDIANS FOR APRIL TESTING — NUMBER OF EXAMPLES
CORRECT

	GRADE				
	IV	V	VI	VII	VIII
PART I					
Approximate number of pupils...	900	480	590	600	600
Test 1.....	7.2	11.6	13.3	12.6	14.0
Test 2.....	2.8	6.2	9.3	8.6	9.2
Test 3.....	1.9	3.6	4.7	4.7	6.4
Test 4.....	.6	1.7	2.9	3.6	4.8
Test 5.....	2.0	3.0	3.7	3.8	4.4
Test 6.....	.4	1.3	2.0	2.3	4.0
PART II					
Approximate number of pupils...	380	760	610	520	460
Test 7.....	2.4	2.7	4.2	4.0	5.1
Test 8.....	.9	2.2	3.7	4.8	5.2
Test 9.....	3.6	5.1	7.5	8.0	9.8
Test 10.....	1.0	1.8	3.8	3.8	5.5
Test 11.....	.4	.8	1.2	1.8	2.4
PART III					
Approximate number of pupils...		370	1000	580	560
Test 12.....		2.0	2.4	2.9	4.6
Test 13.....		1.7	1.6	2.2	3.2
Test 14.....		3.0	5.0	6.5	7.4
Test 15.....		2.3	2.1	2.6	3.7
Test 16.....		2.2	2.7	4.7	6.4
PART IV					
Approximate number of pupils...			440	900	660
Test 17.....			1.4	1.3	2.4
Test 18.....			11.9	11.5	12.9
Test 19.....			2.4	2.1	3.3
Test 20.....			12.5	11.1	13.5
Test 21.....			1.8	1.9	2.5

tion of the field of arithmetic covered by the tests. Any user of the tests is justified in modifying these norms so that they will be in agreement with his objectives.

Limitations. The time allowances are so short that the norms for several of the tests are less than 2.0 examples done correctly. This tends to make the diagnosis unsatisfactory for pupils who work slowly. They do not have a chance to show what they can do. The absence of any tests on the fundamental combinations constitutes a limitation in the intermediate grades and possibly in the case of some pupils in all grades.

6. *Monroe's General Survey Scales in Arithmetic*

General structure. Scale I, designed for Grades III, IV, and V, consists of eight sub-tests. Scale II, designed for Grades VI, VII, and VIII, consists of seven sub-tests. In each sub-test, except No. 7, Scale II, the pupil is asked to do arithmetical examples. In Test 7, he is asked to insert the decimal point in quotients. There are three forms of each scale.

In selecting the sub-tests for each scale, an effort was made to include examples of the types most appropriate for the pupils to whom they would be given. Tests 1, 2, 3, and 4 of Scale I are on the fundamental combinations, or tables, one test being devoted to each operation. These tests are similar to the corresponding tests of the Courtis Standard Research Tests, Series A, and of the Cleveland-Survey Arithmetic Tests. Test 5 calls for single-column addition of five figures. Test 6 consists of subtraction examples, in which the subtrahend is a single figure. Test 7, multiplication, consists of multiplication examples in which the multiplier is a single figure. In Test 8, division, the divisor is a single figure.

The sub-tests of Scale II are represented by the following samples:

46 EDUCATIONAL TESTS AND MEASUREMENTS

ADDITION

Test No. 1

7862	6809	8941	5917	6772	7864	1249
5013	7623	7910	4814	6028	7883	8975
1761	5299	9845	9007	6535	8240	9005
5872	6601	8522	6975	2340	9869	1573
<u>3739</u>	<u>3496</u>	<u>1046</u>	<u>1227</u>	<u>2319</u>	<u>6794</u>	<u>3203</u>

MULTIPLICATION

Test No. 2

4857	5718	6942	4065
<u>36</u>	<u>92</u>	<u>58</u>	<u>47</u>

DIVISION

Test No. 3

41)574	79)36893	32)384	58)27608
--------	----------	--------	----------

SUBTRACTION

Test No. 4

739	1852	975	1087	516	962
<u>367</u>	<u>948</u>	<u>906</u>	<u>821</u>	<u>239</u>	<u>325</u>

ADDITION AND SUBTRACTION OF FRACTIONS

Test No. 5

$$\frac{1}{6} + \frac{1}{3} =$$

$$\frac{3}{4} - \frac{2}{5} =$$

$$\frac{1}{6} + \frac{3}{5} =$$

MULTIPLICATION AND DIVISION OF FRACTIONS

Test No. 6

$$\frac{2}{3} \times \frac{3}{4} =$$

$$\frac{4}{7} \div \frac{2}{3} =$$

$$\frac{5}{12} \times \frac{3}{5} =$$

DECIMAL FRACTIONS

Test No. 7

.03)16.2	Ans.: 54	.07)1.82	Ans.: 26	.05).415	Ans.: 83
.06)7.44	Ans.: 124	.08).952	Ans.: 119	.04)87.6	Ans.: 219
.02).144	Ans.: 72	.08)40.8	Ans.: 51	.09)3.42	Ans.: 38

Construction of duplicate forms. In constructing the duplicate forms of these tests, the figures of the examples were rearranged or changed, so that examples identical in gross structure, but leading to different answers, were obtained. In the case of Scale I, Tests 1, 2, 3, and 4 were constructed by securing at random numbers of the same general magnitude as those used in Form 1. In doing this, especially in Tests 1 and 3, many of the examples were repeated either in their identical form or with the position of the two numbers reversed. Test 5 was constructed by rearranging the same numbers actually used in Test 5 of Form 1. This rearrangement was largely, although not entirely so, a process of shifting the position of the example and reversing the order of numbers therein. In Test 6, two methods were used: either the subtrahends and minuends were grouped together differently or the figures in the minuend were reversed. In Test 7, figures in the multiplicand were rearranged and the resultant number grouped with another multiplier. In Test 8, the same divisors were kept and the dividends either slightly increased or decreased so as to leave the quotients still integral numbers.

In Test 1 of Scale II, the columns of figures in each example were shifted with occasional changes to prevent a zero from coming first in any number. In Test 2, the figures of the multiplicand were rearranged and the resultant number grouped with another multiplier. In Test 3, either the dividends were slightly increased or decreased so as to be evenly divisible by the same divisor, or the dividend or divisor was multiplied or divided by two. In Test 4, either the figures of the minuend were arranged differently or new minuends were constructed by grouping together part of the figures from two minuends to form one, and the remaining figures to form another. The subtrahends were either changed by a different arrangement of figures or by adding or subtracting a

small number, such as ten or twenty, or they were not changed at all. Tests 5 and 6 were constructed by a random selection of fractions of the same general magnitude as those of the same test in Form 1. In Test 7, the position of the decimal point in either divisor or dividend or in both was shifted, and the position of the examples was also changed.

In almost all of the tests an occasional change, not covered by the statements above, was made, either because of the necessity of avoiding impossible combinations or because the result by following too closely the procedure laid down seemed undesirable.

Computation of a single score. The number of examples right is taken as the pupil's point score on each of the sub-tests. These sub-tests yield scores which differ widely in magnitude. Arbitrary rules were adopted for weighting these scores so that approximately equal weight would be given to each test. The weighted sum of the scores on the several sub-tests is the pupil's point score. A single general score is, obviously, a composite, not only of the scores of the different sub-tests, but also of the dimensions of rate and accuracy. However, since a single general measure is desired, this does not constitute a serious criticism of the scale.

The tests for Grades III, IV, and V are entirely different from those given in Grades VI, VII, and VIII. This makes the scores from the two sequences of grades incomparable. They have a different zero point, and it is not unreasonable to expect that they would be expressed in terms of a different unit. It is relatively easy to estimate the approximate difference between the zero points. An estimated difference of 22 has been used as a correction to be added to the point scores obtained from the tests for the upper sequence of grades. Experience with these scales indicates that the difference in the magnitude of the units is not sufficiently large to intro-

duce serious inaccuracies when the corrected scores are considered comparable.

Norms. These scales have been very widely used and the grade norms given below may, therefore, be considered to be representative of present achievements of school children. The norms are for the month of October and are based almost wholly upon first trial scores.

GRADE	SCORE
III	10
IV	21
V	35
VI	44
VII	53
VIII	60

Limitations. The scales are more elaborate than the Courtis Standard Supervisory Tests or the Woody-McCall Mixed Fundamentals (see page 34) and for that reason escape some of the limitations mentioned with reference to these two tests. The time limits for several of the tests are short, but this does not constitute such a significant limitation as in the Cleveland-Survey Tests or the Monroe Diagnostic Tests, because we use only the total score. If errors in timing are accidental and small, the effect on the total score will be relatively slight.

7. The Woody Arithmetic Scales

Measurement by means of a scale. The Woody Arithmetic Scales represent a different type of measuring instrument. They are "scales" rather than "tests." They consist of a set of four scales, one for each of the four fundamental operations. The addition scale of Series A ¹ is repro-

¹ There are two series of four scales each. Series B differs from Series A in that it consists of only certain examples taken from Series A. These were selected so that the increase in difficulty is more regular.

(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
21	32	43	23	$25 + 42 =$	100	9	199	2563
33	59	1	25		33	24	194	1387
35	17	2	16		45	12	295	4954
<u>—</u>	<u>—</u>	<u>13</u>	<u>—</u>		201	15	156	2065
					46	19	<u>—</u>	<u>—</u>

(19)	(20)	(21)	(22)	(23)	(24)	(25)
\$8.75	\$12.50	\$8.00	547	$\frac{1}{3} + \frac{1}{3} =$	4.0125	$\frac{3}{8} + \frac{5}{8} + \frac{7}{8} + \frac{1}{8} =$
1.25	16.75	5.75	197		1.5907	
.49	15.75	2.33	685		4.10	
<u>—</u>	<u>—</u>	4.16	678		8.673	
		.94	456			
		6.32	393			
		<u>—</u>	525			
			240			
			152			

(26)	(27)	(28)	(29)	(30)	(31)	(32)
$12\frac{1}{2}$	$\frac{1}{8} + \frac{1}{4} + \frac{1}{3} =$	$\frac{3}{4} + \frac{1}{4} =$	$4\frac{3}{4}$	$2\frac{1}{2}$	113.46	$\frac{3}{4} + \frac{1}{2} + \frac{1}{4} =$
$62\frac{1}{2}$			$2\frac{1}{4}$	$6\frac{3}{8}$	49.6097	
$12\frac{1}{2}$			<u>$5\frac{1}{4}$</u>	<u>$3\frac{3}{4}$</u>	19.9	
<u>$37\frac{1}{2}$</u>					9.87	
					.0086	
					18.253	
					<u>6.04</u>	

(33)	(34)	(35)	(36)	(37)
.49	$\frac{1}{6} + \frac{3}{8} =$	2 ft. 6 in.	2 yr. 5 mo.	$16\frac{1}{3}$
.28		3 ft. 5 in.	3 yr. 6 mo.	$12\frac{1}{8}$
.63		<u>4 ft. 9 in.</u>	4 yr. 9 mo.	$21\frac{1}{2}$
.95			5 yr. 2 mo.	<u>$32\frac{3}{4}$</u>
1.69			<u>6 yr. 7 mo.</u>	
.22				
.33				
.36				
1.01				
.56				
.88				
.75				
.56				
1.10				
.18				
<u>.56</u>				

$$(38) \quad 25.091 + 100.4 + 25 + 98.28 + 19.3614 =$$

It should be noted that a "scale" differs fundamentally from the tests described above. The examples of a scale are not equally difficult. The author states that his "fundamental idea was to derive a series of scales which would indicate the type of problems [examples] and the difficulty of

the problems [examples] that a class can solve correctly.”¹ According to this function the score of a pupil is a statement of the particular examples which he has done correctly. The author proposed that the score of a class be the degree of difficulty of the example which was done correctly by just fifty per cent of its members. Because the procedure is much simpler, the number of examples done correctly has been taken as a pupil's score by most users of the scales. When this is done the class score is the median of the individual scores.

Function. As generally used, the Woody Arithmetic Scales have a general or survey function similar in some respects to that of the Curtis Standard Research Tests, Series B. Separate measures are yielded for each of the operations, but within each operation the function is even more general than that of the Curtis tests. It should also be noted that ability as measured by the Woody scales means “power” to do increasingly difficult examples. The rate at which the examples can be done is not included in this meaning of ability. Thus “ability” when used in connection with the Woody Arithmetic Scales cannot have the same meaning as is attached to the word when used in connection with the Curtis Standard Research Tests, Series B, or other tests of the same type.

The Woody Arithmetic Scales may be used so as to have a diagnostic function within each operation. However, in order to realize this function, it is necessary to tabulate the results for each example. A portion of the score sheet for a typical class is given in Table IV. The examples not listed in the tabulation were not done incorrectly by more than two pupils. An example not attempted is indicated by a dash. An example done incorrectly is indicated by “1.”

¹ Woody, Clifford, *Measurements of Some Achievements in Arithmetic*, p. 1. (Teachers College Contribution to Education, no. 80, 1916.)

By examining the per cent of examples right at the bottom of the table one learns the types of examples on which this class needs instruction.

TABLE IV. SHOWING THE TABULATION OF THE SCORES MADE BY THE INDIVIDUAL PUPILS OF A SIXTH-GRADE CLASS UPON THE WOODY ADDITION SCALE, SERIES A

PUPIL	EXAMPLES																		
	18	20	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
1.....			1			1			1	1	1		1		1				
2.....					1			-	1	1									
3.....			1	1	1	1		1	1		1		1		1			1	-
4.....			1		1										1				
5.....												1		1				1	1
6.....		1										1		1				1	
7.....					1		1			1	1	1				1	1		
8.....			1					1				1		1	1	1	1		1
9.....		1				1		1			1		1		1			1	1
10.....				1	1	1	1	1	1		1	1	1		1		1	1	1
11.....																			-
12.....															1			1	1
13.....			1			1	1	1		1	1	1	1	1	1			1	1
14.....			1	1		1		1	1	1	1	1	1	1	1	-	-	-	-
15.....			1																1
16.....								1			1				1			1	-
17.....	1							1			1				1			1	1
18.....				1		1	1	1	1	1	1	1	1	1	1		1	1	1
19.....							1												-
20.....			1				1			1	1	1		1	1				-
21.....								1		1	1		1		1			1	
22.....	1	1	1	1		1						1			1	1	1		1
23.....	1		1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1
24.....				1			1	1		1	1		1		1			1	
25.....				1			1	1	1	1	1	1	1		1		1	1	1
26.....			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27.....												1							
28.....			1		1			1			1	1	1	1	1		1	1	1
Per cent right	88	88	57	68	75	65	65	43	72	57	36	46	53	61	34	82	67	37	33

Woody makes the following statement with reference to the uses of the scales:

Perhaps the most valuable use of the scales lies in the diagnosing power of the class mistakes. The writer was convinced during the process of scoring these test papers, nearly 20,000 in all, that the mistakes of a class tend to be grouped around some central tendency. The great variety of the problems in these scales, and the fact that the problems in each of the various operations proceed from the simplest to the more difficult problems, aid greatly in the location of the weaknesses of the class. If a large number in a class fail to invert the divisor in the problems in division of fractions, or if a large number in a class fail to locate the decimal point properly in the problems in multiplication of decimal fractions, a teacher should know immediately that these classes need more practice in these particular processes. In like manner, by locating the particular types of problems missed, one should be able to direct the work of a class more intelligently.

Limitations of the Woody Arithmetic Scales.¹ On the basis of such analyses of arithmetical abilities as have been made, it is clear that all types of examples have not been included in these scales. It also appears from Woody's own statement that those which were chosen were selected, not on the basis of their arithmetical importance, but on the basis of their difficulty and the consistency of the performance of pupils in the various grades. Woody's method of selecting examples may be called "statistical" as opposed to the "analytical method" employed by Curtis, Monroe, and other makers of tests. The statistical method largely neglects the subject-matter field in which the test is being constructed and assumes that an example is suitable for use in a test simply because it is done correctly by a gradually increasing per cent of pupils as one proceeds from grade to grade. It rejects as unfit those examples which do not have this char-

¹ These limitations apply to other measuring instruments of the scale type. They, however, apply more forcibly to the field of arithmetic than to that of some other subjects.

acteristic.¹ On the other hand, the analytical method involves a careful analysis of the field of subject-matter in which the test is being constructed to determine the fundamental types of examples which exist.

When the Woody Arithmetic Scales are used for general survey purposes the method of selecting the examples and the absence of numerous types probably does not constitute a serious limitation. It is, however, necessary to bear in mind that they yield measures of "power" and not measures of "skill." Our objectives in arithmetic are not to teach children to do more and more difficult types of examples, but to do fluently the important types of examples. Hence the measures yielded by these scales are somewhat lacking in consistency with our recognized educational objectives. Incidentally it may be noted that these scales are not published in a convenient form. The scales are printed on four separate sheets instead of a single folder.

In view of the manner in which the examples for the scales were chosen, it seems reasonable that certain limitations should be placed upon Woody's claim for the diagnosing power of these scales. It may also be questioned whether one example is sufficient to test adequately the ability of a class to do examples of that type. For instance, the addition combinations as such are represented by only these two, $2+3$ and $3+1$. Certainly no teachers of experience would accept the performance of a class on these two combinations as an index of their acquaintance with the addition tables. At best the diagnosis can be only partial. Finally, it should be remembered that of the characteristics of specific abilities, accuracy only is measured. The time allowed is sufficient for most pupils to complete all of the examples they are

¹ From what we know about the curve of learning it is doubtful whether this basis can be justified. Ability to do does not increase gradually from grade to grade.

able to do. Hence there is no measure of the rate of work. On the other hand, an analysis of the test papers of a pupil or of a class will yield information which will be valuable in planning future instructions, but as we have just indicated this information is subject to certain limitations.

Norms. The grade norms in Table V are taken from a monograph by the author.¹ Scores for several cities will be found in this source.

TABLE V. GRADE NORMS FOR WOODY ARITHMETIC SCALES IN TERMS OF NUMBER OF EXAMPLES DONE CORRECTLY

	GRADES					
	III	IV	V	VI	VII	VIII
Series A						
Addition.....	15.1	18.9	22.9	29.3	32.0	33.4
Subtraction.....	13.0	16.5	20.9	25.3	28.5	30.8
Multiplication.....	7.2	12.4	19.9	26.9	29.5	33.3
Division.....	7.2	11.3	18.3	25.1	27.2	29.2
Series B						
Addition.....	9.5	12.3	13.7	15.3	15.8	17.0
Subtraction.....	7.3	9.3	10.7	12.6	13.7	14.2
Multiplication.....	5.9	10.4	11.6	14.6	15.9	17.2
Division.....	4.6	6.8	8.4	10.6	11.9	13.0

8. Woody-McCall Mixed Fundamentals

This is a single scale constructed from the Woody Arithmetic Scales, Series B, by selecting appropriate examples from each of the four scales of that series. The first example is addition, the second multiplication, the third division, and the fourth subtraction. In the remainder of the scale the different operations do not appear in any regular order, but

¹ Woody, Clifford, *The Woody Arithmetic Scale*. (Teachers College Bulletin, Eleventh Series, no. 19, May 22, 1920.)

there is always a gradually increasing difficulty. Form I contains thirty-five examples. The second form is similar to the first in structure, but exhibits certain differences in types of examples represented and the order of the operations. It contains thirty-four examples.

This scale is designed to be used for survey or general supervisory purposes. In this respect it is similar to the Curtis Standard Supervisory Tests in arithmetic. It is different in that it is a "power" test. The examples appear to have been selected for their difficulty and without much regard to the types which they represent.

9. Peet-Dearborn Progress Tests in Arithmetic

These tests are described by the authors as including "the leading types of problems that a class needs to master." The Intermediate Series is for Grades IV to V, and the Upper Grade Series for Grades VI, VII, and VIII. Each series consists of five tests, one on problems and one on each of the four operations. All of the tests are "scales." The ones on the operations are similar to Woody's Arithmetic Scales, but contain fewer examples. A pupil's score is the sum of the difficulty values of the exercises done correctly. There is no measure of the rate of work. Since no account of the construction of these tests is available, it is not possible to enter into a detailed discussion of them, but doubtless much of what was said with reference to the Woody Arithmetic Scales applies also to these scales.

10. Lippincott-Chapman Arithmetic Fundamentals Test

This test is a part of the Lippincott-Chapman Classroom Products Survey Tests. It is very similar to the Woody-McCall Mixed Fundamentals Test in both structure and function.

III. STANDARDIZED TESTS FOR MEASURING THE ABILITY TO SOLVE PROBLEMS

The process of problem-solving. "Reasoning" as it occurs in the solving of an arithmetical problem involves these steps: (1) A careful reading of the problem including the association of correct arithmetical meanings with the "technical" terms used in stating the problem. (2) Recall of facts and principles suggested by the problem and required for its solution. (3) Formulation of a hypothesis or plan of solution using as data the results of the first two steps. (4) Verification of this plan of solution. This process of reasoning is usually followed by the calculations outlined in the plan of solution. This additional step, however, is not a part of the reasoning process.

Two kinds of words are used in stating arithmetical problems: (1) The descriptive words give the setting of the problem. Only in an indirect way do these affect the solution. (2) The "technical terms" of an arithmetical problem consist of those words and phrases which define quantities and quantitative relationships. Every problem involves at least three quantities, two given and the third to be found. These quantities are related in a definite way. For example, the sum of the two quantities given equals the third, or the third is the quotient of one divided by the other. In problems involving two or more steps there are more than three quantities and the relationships are more complex. However, in every case there are words or phrases which either directly or indirectly tell what these relationships are, and, consequently, what operations must be performed to obtain the desired answer.

This principle may be illustrated by the following problems: "What are the average daily earnings of a boy who receives \$0.88, \$0.25, \$1.15, \$0.75, \$0.50, and \$0.60 in one week?"

The phrase "average daily earnings" names the quantity to be found and also specifies its relationship with the given quantities. The "average" is the quotient of the sum of the several amounts divided by the number of items. A knowledge of this definite meaning of "average" is necessary if one is formulating a rational plan of solving the problem. If the phrase "average daily" was omitted we should have an entirely different problem.

"How many square yards of linoleum will be required to cover a floor 16 feet by 12 feet?"

"How many square yards" names the third quantity in this problem and in connection with "16 feet by 12 feet" specifies the relations which exist between the quantities. This third quantity is the product of the dimensions divided by nine.¹ In solving this problem the pupil must recall the number of square feet in a square yard and the principle that the area of a rectangle (that is, the figure whose dimensions are given as in the problem) is the product of the length by the width.

In many cases when the first two steps of the reasoning process have been completed satisfactorily, the formulation of the plan of solution (the next step in the reasoning process) involves little uncertainty. In fact it is essentially mechanical. This is the case in these illustrations. In very simple problems, or very familiar problems, the reasoning process is usually short-circuited so that there is no explicit association of meaning with the technical terms nor recall of principles. The problem as a whole or some feature of it serves as a cue for the direct association of the plan of solution. In such cases there is, strictly speaking, no reflective thinking or reasoning, and the mental process involved is much the same as that which occurs in the operations of

¹ An alternative solution is to reduce each dimension to yards before finding the area.

arithmetic. The solution of the problem has become largely automatic.

Specific requirements for a satisfactory reasoning test. The process of problem-solving implies certain requirements for a satisfactory reasoning test. Such a test must give the pupil an adequate opportunity to demonstrate his ability to engage in the reasoning process as it occurs in arithmetic. In order to do this the test must consist of problems which are representative with respect to the reasoning process; that is, with respect to language, facts and principles to be recalled, and complexity. These three factors are interrelated and there is overlapping, particularly between the language of a problem and its complexity, which is determined by the number of quantities involved and the relationships which exist between them.

In order that a pupil's score on a reasoning test may be indicative of his ability to solve arithmetical problems in general, the problems must be carefully selected with reference to content (vocabulary). The ideal reasoning test would be one that included all of the technical terms, but this is not possible because the vocabulary of arithmetical problems is extremely varied and voluminous. In another ¹ place the writer has reproduced twenty-eight different forms of statement which were found in the examination of eight textbooks for the problem, "Given, \$7.50 paid for silk, and price per yard \$1.50, to find the number of yards purchased." This condition makes it necessary to select a few problems which will be representative in respect to content, in order to have a test of usable length.

If we are to secure a measure of "reasoning ability," calculation should be eliminated or reduced to a minimum. This can be accomplished by having the pupil indicate the

¹ Monroe, Walter S., *Measuring the Results of Teaching*, p. 163. (Houghton Mifflin Company, 1918.)

operations to be performed or by disregarding the accuracy of the calculations in computing the pupil's score. The rate of work is relatively less important than in the field of the operations. Excellence in reasoning depends more upon accuracy in formulating an hypothesis than upon doing it rapidly. However, the latter quality is highly desirable.

These requirements only roughly approximated. At present we are able only to approximate the requirements for a satisfactory reasoning test. In the first place, we lack a survey of the language of the problems of arithmetic. Neither do we have adequate statements of the important facts and principles and of the types of complexity. Furthermore, the reasoning process is not mechanical. It tends to be individualistic and adaptable. This adds to the difficulties of the problem of measurement. Hence we must expect to find our present reasoning tests little more than crude measuring instruments.

Function of reasoning tests. All of the reasoning tests described in the following pages are essentially general. They are not diagnostic except for the particular problems included. We should not infer that the errors which a pupil makes on these problems are typical of the errors which he will make in solving other problems. Remedial instruction to correct the errors made in the test is probably desirable, but it does not follow that doing so will make the pupil a good problem-solver in general.

1. Monroe's Standardized Reasoning Tests in Arithmetic

General structure. Test I is designed for Grades IV and V, Test II for Grades VI and VII, and Test III for Grade VIII. Each test consists of fifteen problems printed on a four-page folder so that the pupil has space to work them on the test paper. These problems were selected as being representative of the one- and two-step problems of eight

widely used textbooks.¹ In making this selection special attention was given to language of the problems. In order to make allowance for the variations in difficulty, the amount of credit to be given for each problem has been calculated.² Each problem has two values, "correct principle value," or P, and "correct answer value," or C. These values represent the credit which is to be given for solving the problem correctly in principle and for obtaining the correct answer. In scoring the test papers each problem is marked for correct principle. If a problem is solved correctly in principle, it is further marked with reference to correct answer. A pupil does not receive credit for a correct answer if the problem was solved by the wrong principle. The directions for ad-

¹ For an account of the derivation see Monroe, Walter S., *Report of Division of Educational Tests for 1919-20*. (Bureau of Educational Research Bulletin 5, University of Illinois Bulletin, vol. xviii, no. 21, pp. 36-47.)

² For each problem three records were secured: (1) number of pupils attempting the problem; (2) number of solutions correct in principle; (3) number of correct answers. From these facts the per cent of solutions correct in principle and the per cent of those solved according to the right principle which had also correct answers were calculated. These per cents were translated into sigma values, the former being designated as the "P" value of the problem and the latter as the "C" value. In doing this it was assumed that the ability to solve problems was distributed normally and included between $+2.5$ sigma and -2.5 sigma. In the case of those problems which were solved by the pupils in two successive grades, the average inter-grade interval was found for each group of problems by taking the average of the differences of the sigma values of the problems of the test. This inter-grade interval was added to the values of the problems for the upper of the two grades to reduce them to the basis of the lower grade. The average of the two values was taken as the final value of the problem. The method of weighting is open to criticism. It was used in an attempt to give more credit for doing a difficult problem than for doing an easy one. It is not at all certain that such a plan gives the most truthful indication of a pupil's ability. Some recent studies have shown that unweighted scores correlate very highly with the weighted scores obtained by this method. Therefore, it is likely that the tests would have been nearly as accurate measuring instruments without any determination of weights.

ministering the tests provide for having the pupils mark the problem on which they are working at the end of ten minutes. In this way a rate score may be obtained. It is the sum of the "principle values" of the problems which are solved correctly in principle within ten minutes.

The two forms were constructed so that they were expected to be equivalent. Experience in using them suggests that they are not equivalent, although data are lacking at this time on which a statement concerning their comparability may be based. No attempt was made to construct the different tests so that the scores yielded by the different tests are comparable. Therefore, direct comparisons cannot be made between the fifth- and sixth-grade scores and between the seventh- and eighth-grade scores.

Norms. The number of scores on which the grade norms

TABLE VI. MONROE'S STANDARDIZED REASONING TESTS IN ARITHMETIC — FORM I. — GRADE NORMS FOR APRIL TESTING

	GRADE				
	IV	V	VI	VII	VIII
Correct Principle					
Number of pupils.....	2932	3027	3498	2796	2472
25-percentile.....	6.2	12.1	10.0	13.8	11.5
Median.....	11.3	19.2	14.2	19.7	17.2
75-percentile.....	16.8	25.9	19.4	24.7	22.8
*Rate					
Number of pupils.....	1412	1705	1699	1717	1642
25-percentile.....	5.2	8.0	6.4	8.0	5.3
Median.....	7.8	11.2	8.7	11.2	7.5
75-percentile.....	8.1	15.1	12.1	14.5	10.9
Correct Answers					
Number of pupils.....	2968	2996	3518	2803	2515
25-percentile.....	4.1	7.1	6.9	9.4	5.1
Median.....	7.0	11.3	10.4	13.4	9.0
75-percentile.....	10.7	15.5	14.0	17.4	13.0

* Sum of correct principle values of problems done correctly within ten minutes.

in Table VI are based indicate that they must be considered only tentative. In the opinion of the writer they are too low. In addition to the medians, the 25-percentiles and the 75-percentiles are given. The 25-percentile is the point in the distribution below which there are 25 per cent of the scores. The 75-percentile has a corresponding meaning.

2. *Buckingham's Scale for Problems in Arithmetic*

General structure. The problems for Buckingham's scale were selected largely on the basis of difficulty. In this respect it is similar to Woody's Arithmetic Scales. Division One is for Grades III and IV, Division Two for Grades V and VI, and Division Three for Grades VII and VIII. The problems of Division One increase by steps of approximately 0.3 P.E. from 2.7 to 5.3. The problems of Division Two increase by similar steps of difficulty from 5.5 to 7.3, and the problems of Division Three increase from 7.5 to 9.4. In scoring the test papers attention is given only to the numerical accuracy of the answers. Thus reasoning and calculation are combined in a single score. A pupil's score is the difficulty value of the hardest problem which he answers correctly, unless he has failed on one or more previous problems. In that case, a correction is made by subtracting from the value of the hardest correctly solved problem 0.3 for each failure in Division One, or 0.2 for each failure in Division Two or Three. Thus, if a pupil solved the first six problems in Division One, his score is 4.2; but if he fails on the 4th and 5th (otherwise succeeding through the 6th), his score is 3.6; that is, $4.2 - 2 \times 0.3$.

Although the three divisions of the scale were constructed so that it was expected that the scores obtained from the different divisions would be comparable, the grade medians given in Table VII clearly indicate that the scores are not comparable. The increase in the median score from the

third grade to the fourth grade is 0.8. The increase from the fourth grade to the fifth grade is 1.3. A similar variation is found in the differences between the subsequent grades.

TABLE VII. BUCKINGHAM'S SCALE FOR PROBLEMS IN ARITHMETIC
— FORM I — GRADE NORMS FOR JUNE TESTING

	GRADE					
	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>
No. of pupils.....	4181	4589	7142	5927	6632	5269
25-percentile.....	3.4	4.2	5.7	5.9	7.6	7.7
Median.....	3.8	4.6	5.9	6.4	7.8	8.2
75-percentile.....	4.3	5.2	6.3	6.8	8.3	8.7

Therefore, the scores obtained by the different divisions of the scale are not comparable, although the plan of weighting indicates that they are. The reason for this is that the pupils taking Division Two or Division Three do not have an opportunity to do the problems of the lower divisions. If they did, a number of them would fail to do all of them correctly. Thus, they would receive a score lower than that which they receive when taking only the higher divisions.

In Table VIII the total distributions are given. Evidently a division of the scale higher or lower than that designed for the grade has been used in a few cases. The distributions are significant in that they show that the divisions of the scale are too difficult for the respective grades. The per cent of pupils making zero scores in the third, fifth, seventh, and eighth grades is so large that the scale as now published must be considered unsatisfactory for these grades. This condition could be remedied in the case of Division Two and Division Three by giving the next lower division to the pupils who make zero scores. In the case of Division One, the scale will

TABLE VIII. BUCKINGHAM'S SCALE FOR PROBLEMS IN ARITHMETIC
— FORM I — GRADE DISTRIBUTIONS FOR JUNE TESTING

SCORE	GRADE					
	III	IV	V	VI	VII	VIII
9.0					328	699
8.5				6	782	1084
8.0		1	4	11	1349	1290
7.5			14	13	2931	1740
7.0			240	775	58	44
6.5		2	1012	1886		
6.0		6	1540	1504		
5.5	2	21	3663	1569		
5.0	131	1069	106	57		
4.5	490	1474	14	12		
4.0	815	863				
3.5	1305	798				
3.0	967	255				
2.5	298	75				
0	173	25	549	94	1184	412
Total	4181	4589	7142	5927	6632	5269
Median . . .	3.8	4.6	5.9	6.4	7.8	8.2

have to be extended downward by adding less difficult problems.

3. Other reasoning tests

Some years ago Stone¹ worked out a reasoning test which has been used in several cities, and in a number of city school

¹ Stone, C. W., *Arithmetical Abilities and Some Factors Determining Them*. (Teachers College Contributions to Education, no. 19, 1908.) See also Stone, C. W., *Standardized Reasoning Tests in Arithmetic and How to Utilize Them*. (Teachers College Contributions to Education, no. 83, 1916.)

surveys. Since it was a pioneer test, relatively crude methods of construction were used. It consists of a single list of problems arranged in scale form. It was constructed for use in the sixth grade, but has been used in Grades V to VIII. The time allowance for the test is fifteen minutes. Stone's plan for marking the test papers allows credit for problems partly right and for problems which were not finished. The problem values were determined upon the basis of difficulty.

Courtis included two reasoning tests¹ in his Series A. Starch has devised a test which is called Arithmetical Scale A.² This scale included a number of the problems used by Stone, Courtis, and Thorndike. They were evaluated upon the basis of difficulty and arranged in order of increasing difficulty. The pupils are allowed as much time as they need and a pupil's score is the value of the most difficult problem done correctly. A similar scale is included in the Lippincott-Chapman Classroom Products Survey Tests.

Norms for Stone reasoning test. While this test has been used in many cities, in few have all of the upper grades been tested and the records kept separately by grades. Stone tested in twenty-six different cities, but used only the VI A grade. The test has been applied by others in comparison, but also using only the sixth-grade. The sixth-grade scores for the twenty-six cities tested by Stone give the following results:

Lowest	3.56
Middle	5.50
Highest	9.14

¹ Courtis, S. A., *Manual of Instructions for Giving and Scoring the Courtis Standard Tests in the Three R's*. (Detroit, 1914.) These tests are no longer published.

² Starch, Daniel, "A Scale for Measuring Ability in Arithmetic"; in *Journal of Educational Psychology*, vol. 7, pp. 213-22.

In three cities where school surveys have been made recently the scores were taken separately by grades. The results in these three cities are given in Table IX.

TABLE IX. GRADE NORMS FOR THE STONE REASONING TEST

GRADE	BUTTE, MONTANA*	BRIDGEPORT, CONN.†	SALT LAKE CITY, UTAH‡
V.....	2.2	6.1	3.7
VI.....	3.9	5.2	6.4
VII.....	5.8	6.8	8.6
VIII.....	7.7	4.5	10.5

* *Report of a Survey of the School System of Butte, Montana*, p. 88. (1914.)

† *Report of the Examination of the School System of Bridgeport, Conn.*, p. 102. (1913.)

‡ *Report of a Survey of the School System of Salt Lake City, Utah*, p. 183. (1915.)

Other evidence, drawn from the survey work in these cities, would indicate that the children of Butte were low in ability to reason, the children of Salt Lake City high, and the children of Bridgeport quite uneven.

Recently Stone has issued the following standards:

That 80 per cent or more of 5th grade pupils reach or exceed a score of 5.5 with at least 75 per cent accuracy; that 80 per cent or more of 6th grade pupils reach or exceed a score of 6.5 with at least 80 per cent accuracy; that 80 per cent or more of 7th grade pupils reach or exceed a score of 7.5 with at least 85 per cent accuracy; that 80 per cent or more of 8th grade pupils reach or exceed a score of 8.75 with at least 90 per cent accuracy.¹

IV. INTERPRETATION OF SCORES AND REMEDIAL INSTRUCTION

The scores yielded by standardized tests are useful for several purposes. Among the most important of these are the following: evaluation of the efficiency of the school, classification and promotion of pupils, educational and vocational guidance, and instructional activities. The inter-

¹ Stone, C. W., *Standardized Reasoning Tests in Arithmetic and How to Utilize Them*. (Teachers College Contributions to Education, no. 83, 1916.)

pretation of scores for the first three of these purposes is much the same for the different subjects. Hence we shall reserve our consideration of them until Chapter XIII. The interpretation of scores for instructional activities, and especially the planning of remedial instruction, is not the same for all school subjects. For this reason a section will be devoted to this topic in connection with the discussions of the tests for the different subjects.

The accuracy of individual scores. In interpreting the scores yielded by standardized tests for instructional purposes, it is important that we be informed concerning the degree of accuracy of the individual scores. In the tests we have just considered, precautions have been taken so that in each case most of the sources of error found in written examinations (see pages 5-9) have been eliminated or reduced to a minimum. The plan of marking all examples as either right or wrong insures uniformity, and hence a high degree of objectivity. Several of the tests measure the rate at which the pupil works as well as the quality of his work. By providing each pupil with a printed list of the examples, copying the example either from dictation or from the board is eliminated. It has been shown that pupils do not copy accurately nor do they copy with equal fluency. The elimination of copying the examples eliminates a probable source of error. Objective norms have been provided for the interpretation of the scores.

There are, however, other factors to be considered. A pupil's performance or actual achievement from which his ability is inferred depends upon his physical, mental, and emotional condition. These change from day to day, and from hour to hour, and cause marked variation in successive scores of some pupils. This variation is much greater in some pupils than in others. At any particular time a small per cent of the pupils will be found upon a plane higher than

their normal or average ability, while other pupils will be found at the low ebb of their ability. This fact causes some of the measures to be unreliable in the sense that they are not true indices of the average or normal abilities of certain pupils. However, this happens to a marked degree in only a relatively small per cent of the cases when care is exercised to secure standard conditions in giving the tests, and the number of such cases can be materially reduced by giving the tests a second time and taking the average of the two sets of scores.

Gain or loss in repeating tests. Courtis¹ states that "about one child in ten will have a markedly unreliable score," and "for two thirds of the children the differences will be relatively small." To test the accuracy of the individual scores, and the estimate of Courtis, the writer had the four Courtis tests of Series B given to the pupils in one school a second time. In order that the pupils might not do the same examples, they were asked, on the occasion of the second trial, to begin with the last example and work backward. The results are given in Table X. For addition, the table is read as follows: One pupil did eight fewer examples the second trial, three did four fewer, four did three fewer, five did two fewer, etc. The addition scores average .15 of an example less on the second trial than they did on the first. This difference is indicative of the presence of a constant error in the second-trial scores. Eighty-two per cent of the scores differed by one example, or agreed. Combining the results for the four tests, nearly seventy-three per cent made scores which were the same or differed by only one example. Although this table includes too few cases to warrant any final conclusions, it agrees rather closely with more elaborate investigations.

¹ Courtis, S. A., "Single Measurements with Standard Tests"; in *Elementary School Teacher*, vol. 13, pp. 326-45, 486-504.

TABLE X

SHOWING THE GAIN OF THE SECOND SCORE OVER THE FIRST

(Repetition of Courtis tests, Series B, in one school.)

GAIN	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	TOTAL	AVERAGE GAIN	PER CENT OF SCORES DIFFERING BY ONE EX- AMPLE OR THE SAME
Addition.....	-	-	1	-	-	-	3	4	5	26	37	19	3	1	1	-	100	-.15	82
Subtraction...	-	1	-	-	-	-	1	12	8	22	21	19	10	4	1	-	99	-.10	62
Multiplication	-	-	-	-	-	-	-	2	2	17	41	24	11	2	-	-	99	.16	82
Division.....	1	-	-	-	-	-	1	11	17	22	22	18	4	1	1	-	98	-.28	62
Total.....	1	1	1	-	-	-	5	29	32	87	121	80	28	8	3	-	396	-	-
Per cent.....	.3	.3	.3	-	-	-	1.3	7.3	8.1	22.0	30.6	20.2	7.0	2.0	.8	-	-	-	72.8

The accuracy of individual scores expressed in terms of the coefficient of reliability. The usual method of studying the accuracy of individual scores is to calculate the coefficient of correlation ¹ between the two sets of scores derived from two applications of a test. This is called the coefficient of reliability of the test. These coefficients are not easy to interpret and certain other measures of reliability have been devised. The limitations of space do not permit an explanation of them here, but the writer has published a complete account in another volume.² The reader who is interested in this phase of standardized tests should read this in connection with the present discussion of the accuracy of individual scores. The reliability coefficients for most standardized tests are between .60 and .85 when based upon scores obtained from a single school grade. Only a few above .90 have been reported. For one group of 81 pupils, the addition test of the Courtis Standard Research Tests, Series B, was found to have a reliability coefficient of .87. For other groups of pupils the writer has obtained much lower coefficients. The coefficient of reliability indicates the magnitude of the variable error of measurement.

The accuracy of class scores. The variable errors in individual scores which we have just considered tend to neutralize each other in the computation of scores of classes or larger groups of pupils. However, due to practice effect and other causes, constant errors may be introduced. This frequently happens when a test is given a second time. They do not neutralize each other and affect class scores as well as individual scores. Hence it is very important for one to

¹ See Appendix A for a definition of this and other statistical terms used in this discussion.

² Monroe, Walter S., *An Introduction to the Theory of Educational Measurements*, pp. 188-219; 341-58. (Houghton Mifflin Company, Boston, 1923.)

watch for indications of the presence of constant errors in interpreting the scores of classes.¹

The relation of general intelligence to the interpretation of measures of achievement in school subjects. In the discussion of the interpretation of scores in this chapter and the following chapters, no reference will be made to the need for considering measures of general intelligence in interpreting measures of achievement. A general presentation of this topic will be given in Chapters IX and XII.² However, one should constantly bear in mind that one of the possible causes of a low individual or class score is a low level of intelligence.

Scientific management. During the past few years scientific management has been applied to many forms of human endeavor with results which have been nothing short of marvelous. For example, bricklaying has been practiced by intelligent artisans for centuries, and one might suppose that in the course of that length of time a highly efficient system of laying brick would have been evolved on the basis of accidental successes and imitation, if in no other way. It appears, however, that the method followed has remained practically unchanged for centuries until recently, when the principles of scientific management were applied to the process. A scientific analysis of the process revealed that eighteen motions were made in laying each brick, while only five motions were needed when the material was properly arranged.

Another striking illustration is given by Frederick W.

¹ For a more complete discussion of the accuracy of scores see Monroe, Walter S., *The Constant and Variable Errors of Educational Measurements*. (Bureau of Educational Research Bulletin no. 15, University of Illinois Bulletin, vol. XXI, no. 10, November 5, 1923.)

² See also Monroe, Walter S., *An Introduction to the Theory of Educational Measurements*, pp. 177-80 and 265-69. (Houghton Mifflin Company, Boston, 1923.)

Taylor in his book, *The Principles of Scientific Management*. Some years ago a large quantity of pig iron was being loaded on flat cars at the Bethlehem steel plant. Pig iron is cast in blocks, each of which weighs ninety-two pounds. The method of loading was for the workman to pick up a pig, walk up an incline, deposit it upon the car, walk down, and repeat. The average amount of pig iron loaded per man per day was twelve and one half tons. It was very crude labor, and obviously the amount of pig iron which a workman might load in a day depended upon his physical strength and how he used his strength. If he worked rapidly, with little rest, he soon became exhausted. If he rested too frequently, he wasted his time. A workman's efficiency depended upon his rate of working and the length and distribution of his rest periods. The principles of scientific management were applied to the process, with the result that one workman loaded forty-seven and one half tons a day and the length of the working day was shortened.

These two illustrations are typical of a large number of instances in industrial activities where the efficiency of workmen has been greatly increased by the application of the principles of scientific management. They suggest that the instruction in our schools may be made more efficient by the application of the same principles. For the field of education the principles of scientific management involve (1) an analysis or diagnosis of the teaching situation, and (2) the selection of methods and devices of instruction to meet the situation revealed.

Diagnosis of the teaching situation. In the consideration of the problem of measurement, it was shown that there are as many specific abilities involved in performing the operations of arithmetic as there are types of examples. These operations must be performed with a minimum of attention, so that the focus of the attention may be devoted to the con-

sideration of problems. Thus, the teacher has the problem of engendering in each of her pupils a large number of automatic abilities or specific habits.

The individual differences of pupils furnish another factor of the teaching situation. Pupils differ in native ability and in their past experience. Some pupils are eye-minded, some are ear-minded, and still others are motor-minded. These differences become prominent in their learning. Some pupils grasp quickly the response which is to be made by seeing another perform it, others require a detailed explanation, and still others progress most rapidly by being allowed to reason out the appropriate response. Pupils also differ in the amount of practice they require to reach a given degree of facility in performing an operation.

These two conditions make the teaching situation which the teacher faces a complex one. Before she can intelligently direct her efforts as an instructor she must secure a diagnosis of her class. The tests described in this chapter furnish a means for doing this in the field of arithmetic.

Giving the tests, marking the papers, and tabulating the scores are only first steps. In order to reap the full benefit, the scores obtained from giving a test must be interpreted in terms of pupil needs; and then the instruction must be modified in accordance with the interpretation. Interpretation in terms of pupil needs involves more than simply ascertaining that the class median is above, up to, or below standard, or that certain pupils are above, up to, or below standard. Doing this is equivalent to a physician's telling his patient that his temperature is 103, that his liver is not functioning properly, that he has serious indigestion, and that he is threatened with a nervous breakdown. Clearly this is insufficient. If the physician is to be of service to his patient, he must interpret these facts in terms of his patient's needs. Then, in accordance with these needs, the physician must

prescribe for him — for example, tell him to eat no meat or starchy foods, to spend several hours out of doors each day, and to take certain medicines before each meal and at bedtime. In interpreting scores of educational tests in terms of pupil needs, it is necessary to make similar prescriptions of changes in instruction. The scores of the pupils and of the class, together with the errors, correspond to the temperature, pulse rate, condition of tongue, location of pain, etc., of a patient. They are symptoms; and it is their meanings which are of real importance.

A plan of diagnosis. Different classes as a whole have different needs, and the individual members of a class have a great variety of pupil needs. The first step, however, in interpreting the scores of a class is to ascertain the needs that are common to the class as a whole or to large groups within the class. This should be followed by an individual interpretation of the scores of those pupils at the extremes of the group, particularly of those conspicuously below standard. At present, we do not have sufficient scientific information to permit us to formulate complete statements of the pupil needs which are indicated by the various combinations of symptoms. In the absence of such formulations, it will be helpful to outline a general procedure which may be followed in interpreting the scores of a class after they have been tabulated on the class record sheet. In order that the subject may be as concrete as possible, it will be stated in terms of the Courtis Standard Research Tests, Series B. However, one can easily adapt the general plan to any other test, not only in arithmetic, but also in other subjects.

In a rough way we may group the conditions which will be revealed by this series of tests under six cases as follows:

Case 1. Class medians below standard in both rate and accuracy.

- Case 2. Class medians for rate below standard with satisfactory or high accuracy medians.
- Case 3. Class medians for accuracy below standard with satisfactory or high rate medians.
- Case 4. Scores of members of the class widely scattered.
- Case 5. Irregular development; class medians in one test up to or above standard with class medians in another test distinctly below standard.
- Case 6. Class medians for both rate and accuracy up to or above standard with the individual scores grouped closely about the medians.

Any particular class may exhibit certain combinations of these conditions. For example, the median rate in a test (for example, addition) may be below standard with a satisfactory or high accuracy median and at the same time the scores of the members of the class may be widely scattered. Such a class would be considered as coming under both Case 2 and Case 4.

For each of these cases, it is possible to list certain probable needs. All of these needs may not apply to a given class but some of them will almost certainly be appropriate. In general, it will be profitable to investigate these needs first and determine which of them apply to the class in question. The most probable needs for each case are given below.

Case 1. Class medians below standard in both rate and accuracy. This is probably a case of inefficient teaching. It is true, there may be some other explanation of the low median scores, and it is well, of course, to investigate this possibility. For example, it may be suspected that pupils are below normal in general ability, and evidence on this question may be obtained by giving a test of general intelligence. Or, scores may be below standard because of an unusual course of study, and a comparison of the particular course with a number of successful courses may be made.

Still other causes may be responsible to a greater or less extent for the condition; but inefficient teaching is the most probable single cause.

Case 2. Class medians for rate below standard with satisfactory or high accuracy medians. The most probable need of a class which is below standard in rate with satisfactory accuracy is for *speed drills*. In other words, it is likely that the teacher has not been placing sufficient emphasis upon the rate of work. A possible need is for the elimination of time-consuming methods of work, such as the use of an elaborate phraseology in performing the operations. For example, pupils may be required to name each successive digit in adding a column of figures instead of naming only the partial sums. In some cases slow work is due to the failure of pupils to concentrate attention upon the task. Under such circumstances they should be trained in giving continuous and undivided attention to the task assigned. This includes, among other things, the need for securing a stronger motive.

Case 3. Class medians for accuracy below standard with satisfactory or high rate medians. A low median score in accuracy may be due to the fact that, in working under the timed conditions of the test, the pupils become excited and that in attempting to do a large number of examples they become careless. When this happens, the test should be repeated; and it is likely that, on the second trial, the median for accuracy will be much higher. If, however, the condition persists, the pupils should have frequent timed drills under the conditions of the test and should thus learn to assume the proper attitude toward the test. It may be that the need is even more fundamental. There are many different types of examples in a given operation — for example, in addition. Pupils may have learned to do some types without having learned to do others. For instance, they may be able to add short columns without being able

TABLE XI

SHOWING THE DISTRIBUTION OF THE PUPILS OF A CITY ACCORDING TO THE NUMBER OF EXAMPLES ATTEMPTED, COURTIS STANDARD RESEARCH TESTS, SERIES B

FREQUENCIES IN TERMS OF PER CENTS

Subtraction

GRADE	NUMBER OF EXAMPLES ATTEMPTED																				MEDIAN
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
IV.....	1	7	11	17	19	18	10	8	4	4	1										5.7
V.....		1	3	9	10	18	23	21	8	2	3	2									7.4
VI.....			1	2	8	9	20	16	14	14	9	1	4	2							8.6
VII.....			1		2	6	9	13	12	16	14	12	6	5	2	1			1		10.4
VIII.....				1			9	13	8	19	12	15	8	5	2	3	2	2		1	11.0

to add long columns because of the increased span of attention required. It may be that the pupils have never been given instruction upon the particular type of example occurring in the test. If so, they need to be taught how to do this type.

Case 4. Scores of members of the class widely scattered. The distribution of scores for a typical class will exhibit considerable variability. A few pupils will make very low scores and a few others will make relatively high ones. The distributions shown in Table XI are typical. The scores of many classes exhibit a greater degree of variability. When the scores of the several members of the class are widely scattered, there is need for individual instruction. An analysis of the performance of a pupil is required in order to know his particular needs; but in general, the probable needs of individual pupils are the same as the probable needs of a class. The conditions occurring in Case 4 may often be met by promoting certain pupils to a higher grade and demoting others to a lower grade, provided the general organization of the school will permit such action. When this cannot be done, the teacher must devise some way to give to each pupil that instruction which will meet his needs. Practice exercises, such as the Courtis Standard Practice Tests, have been designed to assist in meeting individual needs. (See page 87.)

Case 5. Irregular development. Irregular development is made possible by the fact that there are a number of different types of examples. Pupils may and do learn to do certain types without learning to do other types with anything like the same rate and accuracy. When irregular development is found to exist, the need is for a redistribution of emphasis in the instruction. Clearly, types for which the scores are up to or above standard need less attention. The Courtis Standard Research Tests, Series B,

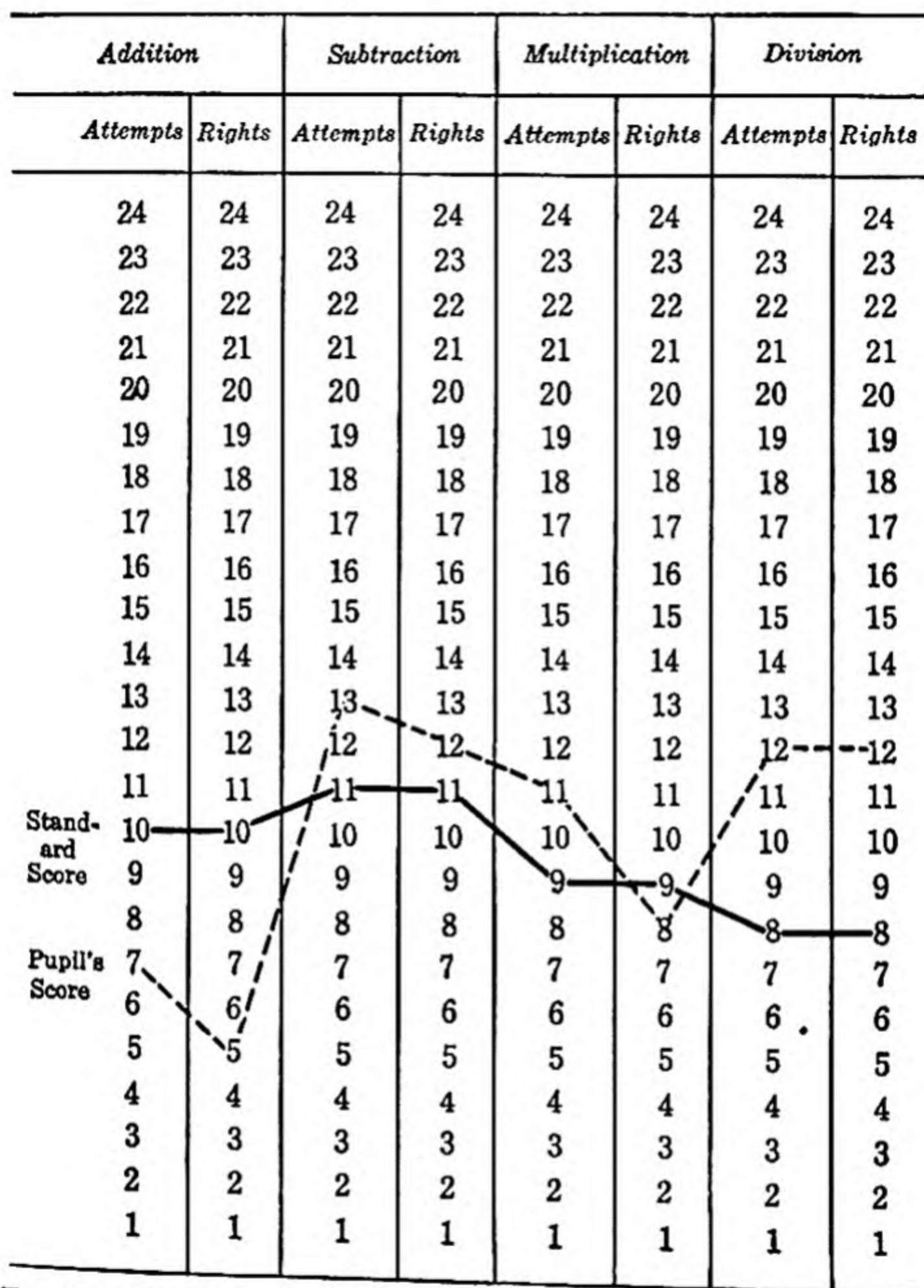


FIG. 3. A CHART, SHOWING THE SCORES MADE BY A SIXTH-GRADE PUPIL, IN COMPARISON WITH THE STANDARD SCORES, USING THE COURTIS ARITHMETIC TESTS

can only show irregularities between the four fundamental operations. The dotted line of Fig. 3 gives the record of a sixth-grade pupil on these tests. The heavy line represents the norms. This girl is weak in addition, but above stand-

ard in subtraction and division. A series of diagnostic tests will reveal irregularities between different types of examples within the same operation.¹

Case 6. Class medians for both rate and accuracy up to or above standard with the individual scores grouped closely about the medians. When a class is above standard in all respects, it may be thought that conditions are satisfactory and that the pupils have no needs. This, however, may not be true. Among the most probable needs are: (1) promotion to the next grade; (2) less time to this subject and more to others, (3) opportunity to take up more advanced topics. The meeting of these needs (especially the one involving promotion to the next grade) depends on the general plan of organization of the school. When a class is just up to standard, pupils may need a continuation of the present instruction or they may now be ready to go on to something else.

Locating a class under one of these cases and ascertaining which of the suggested probable needs apply are the first steps in interpretation. With the partial exception of Case 4, it is mass interpretation and for that reason it is necessarily crude. Like mass instruction, it will not fit all pupils. While the members of the class probably have some common needs, there are also likely to be many individual needs. Thus such mass interpretation should be supplemented by an interpretation of the scores of individual pupils, especially of those having the highest and lowest scores.

In both class interpretation and individual interpretation, it is necessary to remember that the teacher has access, not only to the scores, but also to the test papers which give the performances of the pupils in detail. An analysis of these is helpful, especially where there is any uncertainty about the needs of the pupils. A score reveals nothing as to what was

¹ See Monroe, Walter S., *Measuring Results of Teaching*, pp. 128-35. (Houghton Mifflin Company, Boston, 1918.)

lacking in the mental processes of the pupil, but an analysis of his errors may reveal the defect and suggest corrective instruction. In making such analytical diagnosis, it will be helpful to classify the errors which the pupil has made. In multiplication, the error may be one of multiplication, of placing the partial products, or of addition. In addition, and to a less extent in the other operations, it is frequently impossible to determine the nature of the error. In such cases, one should observe the pupil as he works or even have him "express orally his mental processes." Whenever such detailed studies of the performances of pupils have been made, significant facts have been revealed. This kind of diagnosis requires much time, but the time is profitably employed.¹

Meeting the situation: Laws. In meeting the situation revealed in the case of the operations of arithmetic, the first prerequisite is that the general method of instruction be one suited to engendering specific habits or automatic responses. The laws governing the engendering of specific habits have been quite definitely established.

Stated in psychological terms, *the first law* is that in the beginning the attention of the learner shall be focalized upon the habit to be acquired. In terms of schoolroom practice this means that the learner shall understand what reaction is to be made to a given stimulus, and shall then react to it in the appropriate manner. This gives the learner the right start.

The second law is that the accomplishment of the step outlined in the first law shall be followed by attentive repetitions. It is not sufficient that there be simply repetitions or drill. The drill must be attentive. In the case of the operations of arithmetic this drill may be detached from the solving

¹ See Monroe, Walter S., *Measuring Results of Teaching*, pp. 138-52. (Houghton Mifflin Company, Boston, 1918.)

of problems, or it may be given in the solving of problems.

The third law states that no exception shall be permitted until the habit is firmly established, which means that the attentive practice must be continued until the operation has become a habit, that is, has been made automatic.

The instruction based upon these laws must be adapted not only to the needs of the class, but also to the needs of the individual pupils which the tests reveal. The class needs can be met by placing emphasis upon the types of examples which the pupils as a group are unable to do with standard ability. This emphasis may mean simply more drill, or it may be that the difficulty is due to the pupils' not understanding how the operation is to be performed. If the latter is the case, explanation, or illustration, or opportunity to think it through is needed.

In order to be most effective, the repetitions must be attentive. This means that the drill must be effectively motivated. Arithmetic is one of the best liked of the school subjects. This is particularly true of the operations. This being the case, the motivation of drill in arithmetic is a comparatively simple matter, and in most cases it will be sufficient simply to start the pupils to work and to keep the work from lagging. When more than this is necessary the teacher must demonstrate her resourcefulness by providing an effective method or device for the motivation of arithmetical drill. In the lower grades the playing of certain games provides practice upon certain types of examples. In the upper grades ciphering matches, or, better, the setting of definite standards in both rate and accuracy, are very effective motives.

Individual vs. class needs. However, classes are composed of individual pupils who differ in their needs. Only a few of their needs will be common to the class as a whole.

The usual class instruction in arithmetic does not meet these needs. Frequently the writer has visited classes in arithmetic which were being drilled upon the fundamental operations. A fairly uniform procedure was followed. The same example was dictated to all of the pupils, regardless of whether they needed drill upon this particular type of example or not. Naturally some pupils finished very quickly, and, as they waited for their classmates to finish, there was a tendency for them to become disorderly — a perfectly natural tendency. When a majority of the class had finished the example the teacher stopped the work and read the correct answer. The process was then repeated. The result was that those pupils who worked slowly completed few, if any, examples during the entire period, and, therefore, received little satisfactory drill. The bright pupils spent a considerable proportion of their time waiting on the other members of the class, and probably did not need the particular kind of drill which they received. Obviously, it is the pupil who performs the operation slowly and with difficulty who needs practice. Our present procedure provides drill for the pupil who does not need it, and prevents the pupil who does need it from receiving it in a satisfactory manner.

Modifying the class drill. The type of class instruction described above can easily be modified so as to insure that the slow-working pupils will get some satisfactory drill. Instead of dictating only one example at a time, the teacher can dictate several, and stop the work as soon as a few of the faster workers have finished. The slow-working pupils will have some examples completed.

The teacher must recognize that the rate at which the pupil performs the operations is important, as well as the accuracy. This means that in teaching the teacher must obtain a measure of the pupil's speed, as well as a measure of

his accuracy. If examples are dictated in groups, and the work stopped as suggested in the above paragraph, the number of examples which the pupil does during the class period is a measure of his rate of working. The per cent correct is a measure of his accuracy.

The instruction can be made still more effective if the teacher will prepare a number of sets of examples, each set being confined to examples of the same type. These sets of examples should be written on cards. Then, instead of dictating examples, the teacher can distribute the cards and have the pupils copy the examples from the cards. If the teacher studies the needs of her pupils, it will be possible for her to distribute the cards so that each pupil will have the type of example upon which he needs practice. The pupil is probably injured by being required to practice upon the wrong type of example and, hence, it is very important that each pupil be given the type of example upon which he needs practice.

In one experiment¹ three types of drill were used:

(1) Class drill supplemented by individual assistance on the points of weakness as diagnosed by the results of the test; (2) class drill with extra drill periods provided for the slow pupils, who were drilled in groups rather than individually; and (3) merely class drill with explanations to the class as a whole.

After these types of drill have been used for a month and the results carefully measured, the following conclusions were reached:

(1) All three types of drill produced very large increases in the achievements of the pupils. (2) Class drill supplemented by individual help at the points of weakness as diagnosed by the first test proved much more efficient on account of the exceptional decrease in the variation among the members of the class. This decrease

¹ Smith, James H., "Individual Variations in Arithmetic"; in *Elementary School Journal*, vol. 17, p. 195.

in variation was shown by the decrease in the quartile coefficient of deviation. (3) It has been shown in both the first and second types of drill that individual variations which some writers ascribe to hereditary influences may be greatly modified by appropriate instruction.

Use of practice tests. Courtis has devised a set of Standard Practice Tests¹ which automatically diagnoses each pupil and furnishes the practice which he needs to remedy his defects. These tests consist of forty-eight sets of exercises, which "have been designed to cover every known difficulty in the development of ability in the four operations with whole numbers." These tests are arranged so that the pupils begin the series by taking Lesson 13, a test involving all types of examples found in the first twelve lessons.² All pupils who attain standard ability on this test are excused from the first twelve lessons, because they have demonstrated that they do not need the instruction which these lessons provide. As soon as a pupil who did not attain standard ability on Lesson 13 has finished the first twelve lessons, he takes Lesson 13 again to show that he is now up to standard. Lessons 30, 31, and 44 are also test lessons, and are used in the same way.

Each of the lessons is printed upon a card and a copy is furnished to each pupil. The card is placed beneath a sheet of transparent paper and the example is read through the paper, the work being done on the paper. The lessons have been constructed so that the standard length of time required to complete each one is the same. They are also self-scoring. These two features relieve the teacher of the laborious work of scoring the papers, and make it possible for different pupils to be working upon different lessons at the same time.

¹ Full details regarding these tests may be obtained from the publishers, World Book Company, Yonkers, New York, and Chicago, Illinois.

² All lessons except the test lessons are confined to a single type of example.

Thus, when a pupil has demonstrated that he is up to standard on any type of example, he may at once go on to the next lesson. If he is not up to standard on any lesson his work makes the fact obvious, and he can remain upon that lesson until he acquires the necessary ability without interfering in the least with the work of the other members of the class.

Thus, individual progress is provided for, and at the same time the group information is retained. A considerable saving of pupils' time is effected by excusing from drill those pupils who demonstrate that they possess standard ability. These pupils can spend this time upon other work.

These "Standard Practice Tests" also simplify instruction in ungraded schools. The same lessons are used for all pupils in grades four to eight. Only the time allowed differs. Thus all of the pupils in a rural school could be instructed at the same time and each pupil receive the practice which he needed.

Another series of exercises, known as the "Studebaker Economy Practice Exercises," and based upon some of the same general principles, has been devised by J. W. Studebaker, Assistant Superintendent of Schools, Des Moines, Iowa. They are published by Scott, Foresman & Company, New York and Chicago. Other series of practice exercises have been devised, but, so far as the writer has examined them, they are less complete and give less promise of efficient means of instruction.

However, it must not be forgotten that any set of practice exercises are merely teaching devices. It is more important that the teacher explicitly recognize in her thinking that she is instructing a group of pupils who differ widely in native ability, experience, and training, that all do not learn in the same way, and that a limitation should be placed upon training. When she explicitly recognizes these facts, the

resourceful teacher will find many devices which will be helpful in adapting the instruction to the needs of the pupils.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. Which of the tests described in this chapter would you select in order to secure the most helpful diagnosis of the class and of the individual pupils? Why? Which ones would you use for a general survey? Why?
2. How can the tests described in this chapter be used by the teacher to make her instruction more effective?
3. Do you think pupils will welcome definite objective standards and the use of standardized tests? Why?
4. If you are using standardized tests make charts showing class (or individual) scores in comparison with the standards. Some teachers have found it helpful to have such charts hung in the classroom. It is also helpful to bring such charts to the attention of the patrons of the school.
5. Make a chart showing how the pupils of your class compare with other classes of the same grade and with classes of other grades.
6. Suppose a pupil is unable to do satisfactorily certain types of examples. How would you proceed to locate his particular difficulties? If you are teaching arithmetic try out your plan on some of your pupils.
7. What devices do you use to provide each pupil with the training which he needs? What devices are suggested in this chapter? Can you suggest additional ones?
8. Pupils who are excused from drill because they do not need it should spend their time doing profitable things. Suggest a number of assignments which might be made to such pupils. The assignments may be in subjects other than arithmetic if it seems wise, but they should be such as not to interfere with the instruction of the other pupils.
9. How do you know that the methods and devices of instruction which you are now using are the best? How could you find out?
10. How do you know that you are not giving too much time to arithmetic? How could you find out?
11. Is a class score which is conspicuously above standard always a sign of superior teaching? Why?
12. Construct two tests, each being confined to a single type of example. Give both tests to the same pupils under the same conditions. Compare the two sets of scores.
13. Scientific experimentation will be necessary to determine the best plans of grouping pupils for instruction. These plans are worthy of a trial.
 - a. In a building place together for drill those pupils who are most nearly equal in ability as shown by the tests.

- b. Excuse from drill those who have demonstrated that they are above standard.
- c. Have a special "hospital" class for those pupils who have scores materially below standard. A pupil's sentence to the "hospital" would be until he was up to standard.

SELECTED BIBLIOGRAPHY

/ Only a limited number of references are given in this and the following chapters.

Anderson, C. J. "The Use of the Woody Scale for Diagnostic Purposes"; in *Elementary School Journal*, vol. 18, pp. 770-81. (June, 1918.)

Ashbaugh, E. J. *The Arithmetical Skill of Iowa School Children*. University of Iowa Extension Division Bulletin no. 24. (Iowa City: University of Iowa, 1916. 63 pp.)

Baldwin, Bird T. "The Application of the Courtis Tests in Arithmetic to College Students"; in *School and Society*, vol. 1, pp. 569-76. (April 17, 1915.)

Ballou, F. W. "Improving Instruction through Educational Measurement"; in *Educational Administration and Supervision*, vol. 2, p. 354. (June, 1916.)

Counts, George S. *Arithmetic Tests and Studies in Psychology of Arithmetic*. Supplementary Educational Monograph, vol. 1, no. 4. (Chicago: University of Chicago, 1917.)

Courtis, S. A. "Courtis Tests in Arithmetic; Value to Superintendents and Teachers"; in *Fifteenth Yearbook of the National Society for the Study of Education*, part 1, pp. 91-106. (Bloomington, Illinois: Public School Publishing Company, 1916.)

Courtis, S. A. *Measurement of Classroom Products; The Gary Public Schools*. (New York: General Educational Board, 1919.)

Courtis, S. A. "Capacity, Ability and Performance in Relation to Standard Scores in the Four Operations"; in *Courtis Standard Research Tests Bulletin*, no. 4. (Detroit, Michigan: S. A. Courtis.)

Courtis, S. A. *Third, Fourth, and Fifth Annual Accountings, 1913-1916*. (Detroit, Michigan: Department of Coöperative Research, 1916.)

Dawson, Charles D. "Some Results in Using Starch's Arithmetic Reasoning Test"; in *Journal of Educational Research*, vol. 2, pp. 677-78. (October, 1920.)

Finley, George William. *A Comparative Study of Three Diagnostic Arithmetic Tests*. State Teachers College Bulletin, Series 20, no. 4. (Greeley, Colorado: State Teachers College, 1920. 40 pp.)

Gist, A. S. "Errors in the Fundamentals of Arithmetic"; in *School and Society*, vol. 6, pp. 175-77. (August 11, 1917.)

Gregory, Chester Arthur. *The Efficiency of Oregon School Children in the*

Tool Subjects as Shown by Standard Tests. University of Oregon Publications, vol. 1, no. 1. (Eugene: University of Oregon, 1919. 51 pp.)

Haggerty, M. E. *Arithmetic: A Coöperative Study in Educational Measurements.* University of Indiana Studies, no. 27. (Bloomington: University of Indiana.)

Haggerty, M. E. *Second Report on the Measurement of Arithmetic in Indiana Schools, I.* Indiana University Studies, vol. 3, Study no. 32, pp. 1-58. (Bloomington: Indiana University, 1916.)

Hanus, Paul H., and Gaylor, Harry D. "Courtis Arithmetic Tests Applied to Employees in Business Houses"; in *Educational Administration and Supervision*, vol. 3, pp. 505-21. (November, 1917.)

Heckert, J. W. "The Cleveland Survey Tests in Arithmetic in the Miami Valley"; in *Elementary School Journal*, vol. 18, pp. 447-57. (February, 1918.)

Judd, Charles H. *Measuring the Work of the Public Schools.* Cleveland Education Survey. (Cleveland, Ohio: The Survey Committee of the Cleveland Foundation, 1916. 290 pp.)

Kallom, Arthur W. "Analysis of and Testing in Common Fractions"; in *Journal of Educational Research*, vol. 3, pp. 177-92. (March, 1920.)

Keener, Edward E. "The Cleveland-Survey Arithmetic Test in Grade 5 B in Chicago"; in *Chicago Schools Journal*, vol. 4, pp. 336-44. (May, 1922.)

Mead, Cyrus D., and Johnson, Charles W. "Testing Practice Material in the Fundamentals"; in *Journal of Educational Psychology*, vol. 9, pp. 287-97. (May, 1918.)

Monroe, Walter S. *A Report of the Use of the Courtis Standard Research Tests in Arithmetic in Twenty-Four Cities.* Bureau of Educational Measurements and Standards Studies no. 4. (Emporia: Kansas State Normal School, 1915. 94 pp.)

Monroe, Walter S. *The Illinois Examination.* University of Illinois Bulletin, vol. 19, no. 9, Bureau of Educational Research Bulletin, no. 5. (Urbana: University of Illinois, 1921. 70 pp.)

Monroe, Walter S. (edited by). *Studies in Arithmetic, 1916-17.* University of Indiana Studies, vol. 5, Bureau of Coöperative Research Study no. 38. (Bloomington: University of Indiana, 1918. 40 pp.)

Monroe, Walter S. "Derivation of Reasoning Tests in Arithmetic"; in *School and Society*, vol. 8, pp. 295-99, 324-29. (September 7 and September 14, 1918.)

Monroe, Walter S. "A Series of Diagnostic Tests in Arithmetic"; in *Elementary School Journal*, vol. 19, pp. 585-607. (April, 1919.)

Monroe, Walter S. "The Ability to Place the Decimal Point in Division"; in *Elementary School Journal*, vol. 18, pp. 287-93. (December, 1917.)

Morrison, J. Cayce. "The Supervisor's Use of Standard Tests of Efficiency"; in *Elementary School Journal*, vol. 17, pp. 335-54. (January, 1917.)

92 EDUCATIONAL TESTS AND MEASUREMENTS

Osburn, W. J. *Diagnostic and Remedial Treatment for Errors in Arithmetical Reasoning*. (Madison, Wisconsin: State Department of Public Instruction, 1922. 12 pp.)

Otis, Arthur S., and Davidson, Percy E. "The Reliability of Standard Scores in Adding Ability"; in *Elementary School Teacher*, vol. 13, pp. 91-105. (October, 1912.)

Peterson, Joseph. "Methods of Interpreting Results in the Cleveland Arithmetic Tests"; in *Journal of Educational Research*, vol. 3, pp. 280-92. (April, 1921.)

Scott, Colin A. "An Eighth-Grade Demonstration Class and the Three R's"; in *Journal of Educational Psychology*, vol. 10, pp. 189-218. (April, 1919.)

Sexton, Elmer K. *Arithmetic Study in the Public Schools of Newark, New Jersey*. Board of Education Monograph. (Newark, N.J.: Board of Education, 1919. 30 pp.)

Smith, James H. "Individual Variations in Arithmetic"; in *Elementary School Journal*, vol. 17; pp. 195-200. (November, 1916.)

Spaulding, F. E. *The Arithmetical Abilities of School Children as Shown by Courtis Tests*. Board of Education Bulletin no. 1. (Cleveland: Board of Education, Division of Reference and Research, 1917. 15 pp.)

Starch, Daniel. "A Scale for Measuring Ability in Arithmetic"; in *Journal of Educational Psychology*, vol. 7, pp. 213-22. (April, 1916.)

Stone, C. W. *Arithmetical Abilities and Some Factors Determining Them*. Teachers College Contributions to Education no. 19. (New York: Teachers College.)

Swift, G. C. "Standard Tests for Teachers' Use"; in *School and Society*, vol. 8, pp. 117-18. (July 27, 1918.)

Taylor, E. H. "A Comparison of the Arithmetical Abilities of Rural and City School Children"; in *Journal of Educational Psychology*, vol. 5, pp. 461-66. (October, 1914.)

Theisen, W. W., and Fleming, Cecile White. "The Diagnostic Value of the Woody Arithmetic Scales; a Reply. Part I"; in *Journal of Educational Psychology*, vol. 9, pp. 475-88. (November, 1918.)

Uhl, W. L. "The Use of Standardized Materials in Arithmetic for Diagnosing Pupils' Methods of Work"; in *Elementary School Journal*, vol. 18, pp. 215-18. (November, 1917.)

Wilber, Flora. *Experiments with Courtis Practice Pads*. University of Indiana Studies, vol. 3, no. 32, pp. 103-10. (Bloomington: University of Indiana, 1916.)

Willing, Matthew H. "The Encouragement of Individual Instruction by Means of Standardized Tests"; in *Journal of Educational Research*, vol. 3, pp. 193-98. (March, 1920.)

Wilson, Estaline. "Improving the Ability to Read Arithmetic Problems"; in *Elementary School Journal*, vol. 22, pp. 380-86. (January, 1922.)

Wilson, G. M. "The Proper Content of a Standard Test"; in *Elementary School Journal*, vol. 19, pp. 375-81. (January, 1919.)

Wood, Ernest R. "Tests of Efficiency in Arithmetic"; in *Elementary School Journal*, vol. 17, pp. 446-53. (February, 1917.)

Woody, Clifford. *Measurements of Some Achievements in Arithmetic*. Teachers College Contributions to Education, no. 80. (New York: Teachers College.)

Zeidler, R. "Tests of Efficiency in the Rural and Village Schools of Santa Clara, California"; in *Elementary School Journal*, vol. 16, pp. 542-55. (June, 1916.)

CHAPTER III

READING

I. SILENT READING

The complex nature of silent reading. Any discussion of the measurement of reading ability naturally falls into two major divisions — silent reading and oral reading. The activity of silent reading is almost wholly confined to “comprehending sentences and paragraphs seen.” “Understanding of words seen” is included in this, but we are seldom concerned with it as a separate activity. The comprehension of sentences and paragraphs seen includes a number of mechanical elements, such as eye-movement and fixation. These may be said to form the mechanism of silent reading. The functioning of this mechanism is accompanied by the mental processes which are required in associating meaning with the words, sentences, and paragraphs seen. This activity is assimilative as well as associative. The elements and meanings which are associated with words and phrases are combined, giving an appropriate weight to each element. The resulting ideas may be further combined in major units and with ideas already possessed by the reader.

The degree of meaning associated with words and sentences varies from the very meager meaning required for a general comprehension, or merely grasping the trend of the story, to the rich meaning required for the complete and precise understanding of the passage read. The degree of meaning depends upon the reader's purpose, or the “set” of his mind. It is also affected by the nature of the material read. Narratives in which the action of the story stands out clearly are likely to be read for the story instead of studied.

In fact, the simple narratives found in our school readers for primary grades offer little challenge for study. On the other hand, philosophic writings present a challenge for intensive study. It is necessary to associate concise meanings with the words and phrases used in these writings. In order to get the full significance of the passage, it is frequently necessary to compare the ideas directly associated with it with ideas from other sources. It may also be noted that poetry offers a challenge to the reader which is different from that presented by prose. A textbook is likely to be read in a way different from that in which a light novel is read.

The traits to be measured. In the teaching of silent reading and in its use in the study of other school subjects, the central objective is suggested by the query, "How much did you get out of it?" "The amount of meaning" which a pupil acquires within a given time depends upon two things — his rate of reading and his degree of comprehension. A pupil who reads slowly and with a high degree of comprehension may obtain more ideas and thus get more out of what he reads within a given time than another pupil who reads rapidly, but with a less degree of comprehension. Another way of expressing this concept is that a pupil's achievement in silent reading is two-dimensional, and a complete description of it must include a measure of both his rate of reading and the degree of his comprehension. For certain purposes we may require the measurement of only one of these traits, but in general they should be measured separately, or, if the test yields only a single score, it should depend upon both of these dimensions.

Courtis has stated that "the rate of reading in itself is probably not a significant measure of anything." This statement is misleading. The pupil's rate of reading is a symptom of his reading mechanism. A pupil who has not acquired a good reading mechanism cannot read rapidly.

The relation between the rate of his reading and the quality of his reading mechanism is, however, not a fixed one. A pupil who has a good mechanism may not read to his fullest capacity. Unfamiliar words may greatly reduce his rate. His rate may also be lowered because he is engaged in a type of reading which requires a high degree of comprehension and assimilation; or his rate may be very high because he is skimming the selection with a very limited understanding of it. With these limits, however, a pupil's rate of reading may be considered a reliable symptom of his reading mechanism.

Types of reading. In arithmetic we recognized a number of types of examples which called for essentially different abilities. As yet our analysis of the field of silent reading is crude, but it appears likely that there are a number of different types or kinds of silent reading. We have one type when a pupil is reading for the purpose of memorizing. If he is searching for information bearing on a particular question or problem, the outcome will be different. If his purpose is to secure general information there will be a still different outcome. As we have pointed out above, one potent factor in determining the type of reading is the nature of the material. Thus a pupil's scores for rate and comprehension should be stated with reference to both his purpose and the kind of material read — poetry, prose, narrative, description, etc. A pupil's rate of reading and degree of comprehension are also influenced by the difficulty of the material.

The problem of measurement. The problem of measurement for ability in silent reading is more difficult than that for arithmetic. The functioning of arithmetical ability results in an outcome which may be accurately observed.¹ In

¹ It may be urged that the outcome of the functioning of arithmetical ability is essentially mental and that the answers spoken orally or expressed in written form should not be considered the true outcome. If this position

silent reading the rate may be observed directly by noting the elapsed time for the reading of a given passage. Comprehension is mental and cannot be observed. Frequently a pupil's bodily posture and facial expression may give some indication of his comprehension, but it is clear that they cannot be used as a basis of a comprehension score. To secure a satisfactory basis for estimating the degree of a pupil's comprehension, it is necessary to require of him a supplementary performance which can be accurately observed. For group testing it is necessary that this be a written performance. It should be one directly dependent on the mental outcomes of the silent reading process and also one which makes minimum demands upon other abilities.

In the teaching of silent reading and in the use of it in the study of other school subjects, the performances most frequently required of pupils are to tell what they have read, or to answer questions about it. Experience with silent reading tests has demonstrated that one which requires a pupil to reproduce from memory what he has read is unsatisfactory for measuring silent reading ability. It is only slightly more satisfactory to have him answer questions of the usual type. The most significant objection to tests requiring such performances is that the scoring is highly subjective. Competent persons tend to disagree in assigning scores to the same test papers.

In order that the measurements of silent reading ability may be as accurate as possible, it is necessary to require a type of performance which will be marked the same by different persons. Test-makers have exhibited much ingenuity in the invention of exercises which would meet this require-

is taken, it may be pointed out that the expression of the answer is a normal supplementary activity that is done without focusing the attention upon it. The writer, however, prefers to consider the expression of the answer as a direct outcome of the functioning of arithmetical ability.

ment. In the most widely used silent reading tests the pupil is asked to answer a question with the test before him, but the question is one for which only one answer can be considered correct. In most cases the test is constructed so that the pupil does very little writing in answering the questions. Several typical exercises will be reproduced as a part of the descriptions of particular silent reading tests.

General limitations of silent reading tests. The fact that, in measuring ability to read silently, it is necessary to require a supplementary performance in order to secure one that can be accurately observed, imposes a significant limitation upon the measures yielded by silent reading tests. A pupil's score is a measure of his silent reading ability plus his ability to do something else. Furthermore, the exercises of the tests tend to create artificial reading situations. In many of the tests the pupil does not read continuous material. For these reasons some have expressed doubt concerning the validity of the measures yielded by our present silent reading tests. Since we are able to learn about the degree of a pupil's comprehension only through some supplementary performance, we are handicapped in demonstrating that silent reading tests really measure silent reading ability. It, however, appears reasonably certain that for practical purposes we are justified in assuming that they do yield measures of certain types of silent reading ability.

If we are correct in our thesis that there are several types of silent reading, it is necessary to have in mind that a single test probably measures only one type of silent reading ability. In order to secure diagnostic measurements similar to those yielded by diagnostic tests in the field of arithmetic, it would be necessary to use a battery of different types of silent reading tests. A survey of the existing tests indicates that we have satisfactory ones for only a few of the types of silent reading. Hence, it is not now possible to secure a

complete diagnosis of pupils with respect to ability to read silently.

Silent reading tests are lacking in reliability. When the test is repeated, many pupils will make different scores. The situation is about the same as for arithmetic tests. (See page 69.) The scoring of certain tests is somewhat subjective, but the most potent cause of the unreliability is the variability of the performances of the pupils. In addition, acquaintance with the nature of the exercises tends to increase the scores of pupils even when a different form of the test is used. In this way a constant error is introduced in the scores when a test is given a second time.

1. Monroe's Standardized Silent Reading Tests

General structure. Test I is for Grades III, IV, and V, and Test II for Grades VI, VII, and VIII, and Test III for the high school. There are three forms of Test I and Test II, and two of Test III. Each test consists of a series of exercises. Each exercise consists of a short paragraph to be read and a single question based on it. These tests are very similar to the Kansas Silent Reading Tests devised by F. J. Kelly. In fact, they are essentially a revision of these tests which were the first of this type to be constructed.

Most of the paragraphs on which the exercises are based were taken from school readers and other books which children read. In the construction of the tests a large number of exercises were prepared. These were given to a number of pupils in such a way that the following facts could be ascertained for each exercise: (1) the total number of pupils taking it; (2) the number who did it correctly; and (3) the total time used by these pupils. A "comprehension value" was calculated from these three items of information.¹ A "rate

¹ For the details of the construction of these tests see Monroe, Walter S., "Monroe's Standardized Silent Reading Tests"; in *Journal of Educational Psychology*, vol. 9, pp. 303-12. (June, 1918.)

of value" was calculated by dividing the number of words in the exercise by 5. Some of the exercises were found to be ambiguous. Others were not highly objective in scoring. A few were too easy and others appeared to be too difficult. In constructing the final tests all exercises were rejected which seemed to be unsatisfactory for testing purposes. The exercises are arranged in order of ascending difficulty, but the tests are not "scales" in the sense that Woody's Arithmetic Scales are "scales."

Explanation of the test to pupils. In their regular school work pupils are not accustomed to exercises of the type found in these tests. Hence it is necessary to make them acquainted with the sort of thing they are to do before they begin the test. This is accomplished by means of a sample exercise and specific directions. These are reproduced below.

INSTRUCTIONS TO BE READ BY TEACHER AND PUPILS TOGETHER

This brief test is given to see how quickly and accurately pupils can read silently. To show what sort of test it is, let us read this:

I am a little dark-skinned girl. I wear a slip of brown buckskin and a pair of soft moccasins. I live in a wigwam. What kind of a girl do you think I am?

Chinese French Indian African Eskimo

The answer to this exercise is "Indian," and it is to be indicated by drawing a line under the word. The test consists of a number of exercises like this one. In some of the exercises you are told to draw a line under the word which is the right answer, or to mark it in some other way, and in some you are to write out your answer. If an exercise is wrong it will not count, so it is wise to study each one carefully until you know exactly what you are asked to do. The number of exercises which you can finish thus in five minutes will make your score, so do them as fast as you can, being sure to do them right. Stop at once when time is called. Do not open the papers until told, so that all may begin at the same time.

The teacher should then be sure that each pupil has a good pencil or pen. Note the minute and second by the watch and say, **BEGIN**.

ALLOW EXACTLY FIVE MINUTES

Answer no questions of the pupils which arise from not understanding what to do with any given exercise.

When time is up say **STOP** and then collect the papers at once.

The pupil's comprehension score is the sum of the "comprehension values" of those exercises which he does correctly in the time allowed. His rate score is the sum of the "rate values" of all of the exercises which he has done whether correctly or incorrectly. The pupil is supposed to do the exercises in order, and hence an exercise skipped is counted as done incorrectly.

Function. The function of these tests may be defined as to measure the ability of pupils to read simple paragraphs for the purpose of answering specific questions. It is obvious that the scope of the test is general within this type of reading and that only indirectly do they yield measures of other types of reading.

Limitations. A significant limitation of Monroe's Standardized Silent Reading Tests is the narrowness of function just noted. Although our analysis of the field of silent reading is very crude, it appears that there are a number of different types of reading depending upon our purposes and the type and difficulty of the material read. If this is the case, a silent reading test consisting of a single type of exercise cannot yield a comprehensive measure of silent reading ability. Hence scores yielded by Monroe's Standardized Silent Reading Tests should not be used as *general* measures of ability to read silently. They are measures only of the type of silent reading defined by the character of the test.

Experience with these tests has shown that the scoring is not as highly objective as is desirable. Many pupils com-

plete the test before the end of the time allowed. The measures lack precision because there is no way of giving partial credit for the exercise on which a pupil is working when time is called. Some of the exercises have been criticized as being too much like puzzles. Also a pupil is not doing continuous reading. The reading situation is artificial and many pupils fail to obtain from the directions a sufficient understanding of what they are to do. Several of these defects have been corrected in Monroe's Standardized Silent Reading Tests, Revised, but certain of the limitations are inherent in the character of the test. Since the original tests, except Test III, have now been replaced by the revision, a statement of the norms is omitted.

2. *Monroe's Standardized Silent Reading Tests, Revised*

General structure. These tests are very similar to the tests just described. The revision includes only Tests I and II. The pupil gives his answer to an exercise by drawing a line under a word or by indicating it in some other way. No writing is required. In most exercises the pupil has to select one out of five words in making his response; in a few there are only four words. The nature of the exercises may be illustrated by the accompanying sample from Test II.

The exercises for Form 1 were taken with some modifications from the original edition of Monroe's Standardized Silent Reading Tests. In selecting the exercises, and in making the modifications, an effort was made to have all the exercises approximately the same length. Those for Test II are slightly longer than the ones included in Test I. The exercises are not absolutely equal with respect to difficulty. They are arranged so that in general there is a slight increase in difficulty from exercise to exercise, but in no sense can they be considered a difficulty scale. To secure absolute uniformity with respect to difficulty would have required the ex-

1. Locksley shot his arrow as carelessly in appearance as if he had
not even looked at the mark, yet it alighted two inches nearer to the
center than that of Hubert.

Draw a line under the word which best describes Locksley.

skilful strong handsome tall wise

2. It took us an hour to transport our stores to the spot selected
for the encampment. Having pitched our tent, using the five oars to
support the canvas, we got out our lines, and went down to fish.

Draw a line under the word which tells what these people were
doing.

studying camping swimming hunting working

penditure of a prohibitive amount of labor. Even if this were not true, it is believed to be desirable to have a moderate range of difficulty in view of the fact that the exercises are to be given to pupils in a sequence of three successive school grades.

The nature of the exercises is explained to the pupils by means of three fore-exercises instead of one as in the original tests. In addition they are given certain verbal explanations. Four minutes are allowed for the test in all grades. This time allowance was intended to be such that practically no pupils would complete the tests; but it has been found that this is not always the case.

A few of the exercises are based on poetry, although the majority are based on prose paragraphs. In the light of a recent investigation ¹ it appears likely that the reading of poetry, even for the purpose of answering a question, is not the same activity as the reading of prose for the same purpose. To the extent that this is true the test is not consistent, and this constitutes one of its limitations. In a number of other respects a high degree of uniformity has been secured. In every case the pupil makes the same kind of response. The exercises are approximately equal in length and the questions asked appear to call for much the same type of reading. The pupil receives two scores. His comprehension score is the number of exercises which he answers correctly. His rate score is the number of words which he reads per minute, or the total number of words read divided by 4. In order to obtain this rate, the pupil is asked to mark the line which he is reading when the signal to stop is given. The cumulative totals of the words are printed in the left-hand margin so that the total number of words read is easily obtained.

¹ Pressey, L. W., and Pressey, S. L., "A Critical Study of the Concept of Silent Reading Ability"; in *Journal of Educational Psychology*, vol. 12, pp. 25-31. (January, 1921.)

Correction of scores derived from the test designed for Grades VI, VII, and VIII. The test for Grades III, IV, and V is entirely different from the one given in Grades VI, VII, and VIII. This makes the scores from the two sequences of grades incomparable. They have a different zero point, and

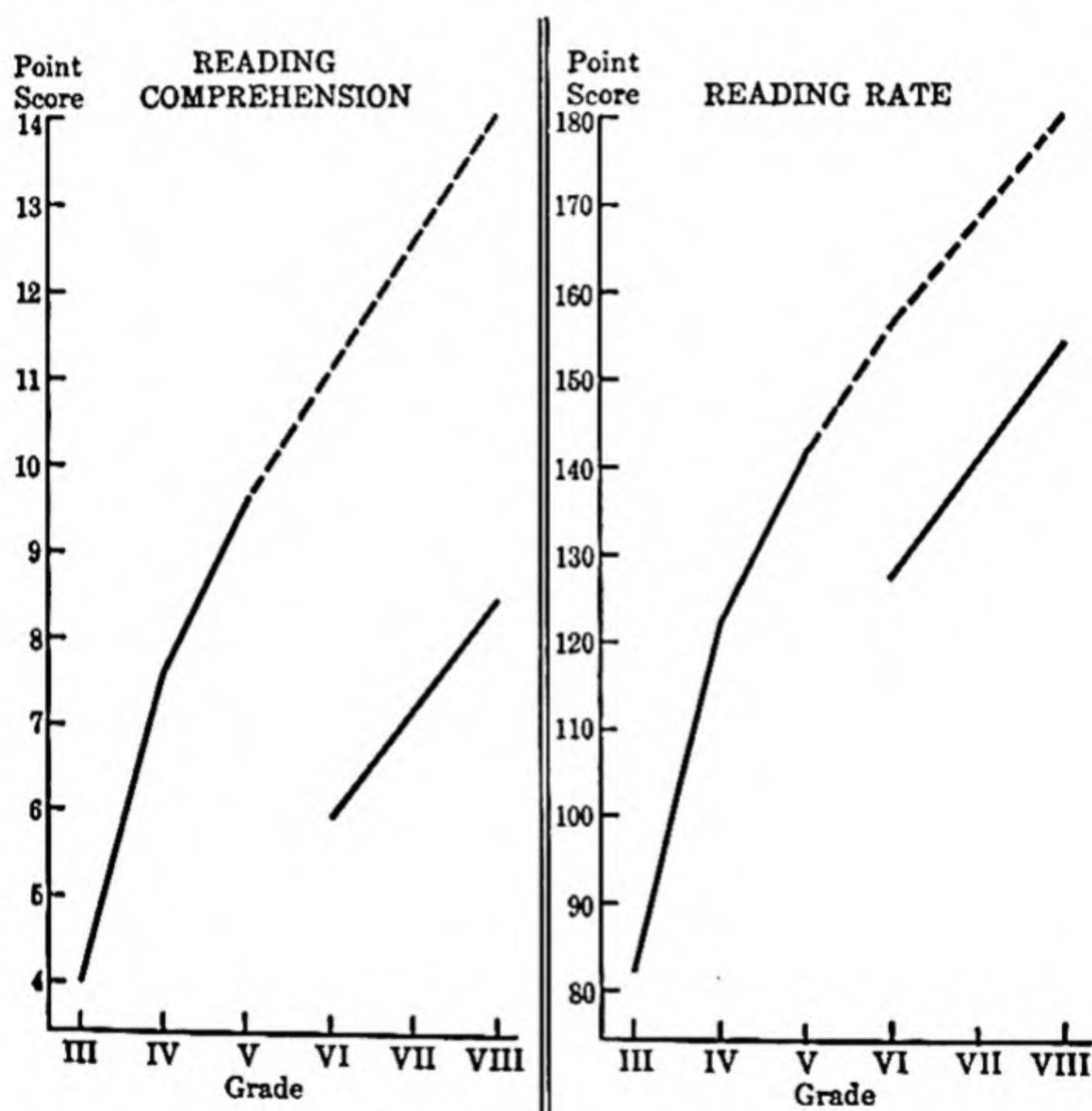


FIG. 4. SHOWING METHOD OF ESTIMATING CORRECTIONS TO BE ADDED TO SCORES OF TEST II

it is not unreasonable to expect that they would be expressed in terms of a different unit. It is relatively easy to estimate the approximate difference between the zero points. This difference can be used as a correction to be added to the point scores obtained from the tests for the upper sequence of grades. This was done, and it appears that the differ-

ences in the units are not sufficiently large to introduce serious inaccuracies when the corrected scores are considered comparable.

The median for each grade was represented graphically, as shown in Fig. 4. The curve of progress for Grades III, IV, and V was then extended so that the extension would parallel the curve of progress for Grades VI, VII, and VIII. This extension, together with the progress curve for the lower sequence of grades, forms a progress curve for Grades III to VIII inclusive. The distance between the extension and the original curve is assumed to represent the difference in the zero points of the two tests.

The estimated corrections are: rate, 29, and comprehension, 5. When these corrections are added to the scores derived from Test II, the scores will, in general, be approximately comparable to the corresponding scores derived from Test I.

Equivalence of duplicate forms. In studying the equivalence of the three forms of Monroe's Standardized Silent Reading Tests, Revised, Forms 1 and 2 were found to be approximately equivalent especially in comprehension. Form 3 is slightly easier and yields higher scores.¹

Norms. These tests have been very widely used and the norms given in Table XII are based upon several thousand scores in each grade. They are for October testing.

3. Curtis Silent Reading Test No. 2

General structure. This test is designed for use in Grades II to VI. There are four forms. It is most satisfactory in the lower grades. The test consists of a single continuous story which the pupil reads for three minutes, but at the end

¹ Monroe, Walter S., *The Illinois Examination*, Bureau of Educational Research, *Bulletin* no. 6, University of Illinois, *Bulletin*, vol. XIX, no. 9, p. 12.

TABLE XII. NORMS FOR MONROE'S STANDARDIZED SILENT READING TESTS, REVISED, FORM 1, OCTOBER SCORES

GRADE	SILENT READING	
	Comprehension	Rate
III.....	3.8	82
IV.....	7.7	122
V.....	9.8	142
VI.....	11.0	158
VII.....	12.1	170
VIII.....	13.5	183

of each half minute a signal is given and he is to mark the last word which he has just read and keep on reading. The general character of the story is illustrated by the accompanying sample (see page 108).

In the second part of the test the pupil is required to answer simple questions on what he has read. The story is repeated so that he may read it again if he finds it necessary. A portion of Part II is reproduced below.

When the day of the party came, Daddy planted a May-pole and Mother tied it with gay-colored ribbons. There were to be games and dances on the grass and a delicious supper, with a basket full of flowers for every child.

1. Were the children to have anything to eat?.....—
2. Were they going to play on the grass?.....—
3. Were they going into the house to dance?.....—
4. Were the baskets to be full of flowers?.....—
5. Was it Daddy who tied the ribbons to the pole?.....—

The Kitten Who Played May-Queen

When the spring sun was warm and the spring flowers had begun to bloom, our children had a May-party out on the lawn. Every little boy or girl who lived nearby was invited. Betty wrote the invitations and Bobby carried them to the children.

When the day of the party came, Daddy planted a May-pole and Mother tied it with gay-colored ribbons. There were to be games and dances on the grass and a delicious supper, with a basket full of flowers for every child.

5
13
21
30
39
47

49
58
66
74
82

Under the big tree in front of the house was a large box. Mother covered it with a rug and put a little chair on it. This was the throne on which the May-queen was to sit.

6. Was the large box under a big tree?.....—
7. Was the big tree back of the house?.....—
8. Was it the queen who put the chair on the rug?.....—
9. Was it Daddy who put the rug on the box?.....—
10. Was the queen to sit on the little chair?.....—

The pupils are given five minutes to answer as many of the questions as they can. The measure of their understanding of the story is expressed in terms of the number of questions answered and the index of comprehension. This index is found as follows: "Subtract the wrong answers from the right answers. (If there are more wrong than right, find the difference and give it a negative sign.) Divide the difference by the number of right answers, carrying the result to three places and keeping two." This calculation is made from a table so that the labor involved is reduced to a minimum. In addition to these two scores, the number of words read per minute is obtained from the first part of the test. This is a measure of the pupil's rate of reading. The scores of a pupil are recorded on an individual record card which is more convenient to handle in recording the scores of the class on the class record sheet.

The three scores, rate of reading (words per minute), number of questions, and the index of comprehension are recorded on the forms marked Table 1 and Table 3 in Fig. 5. In Table 1, 0 means 0 to 19, 20 means 20 to 39, and so on. Table 3 represents a different type of form for recording scores. The test papers, or in this case the individual cards containing the record of the pupils' scores, are first sorted into piles according to the number of questions answered, putting into the first pile all the cards having a score of 0 to 4 questions answered, into the second pile all those cards

Table 1

Rate of
Reading

Score in words per minute	Number of chil- dren making each score
Over	
400	
380	
360	
340	
320	
300	
380	
360	
340	
320	
300	
280	
260	
240	
220	
200	
180	
160	
140	
120	
100	
80	
60	
40	
20	
0	
Total	
Median	

Table 3

Index of Comprehension

Diagnosis		Guesswork			Comprehension poor; additional training needed					Comprehen- sion satis- factory		
Questions Answered	Total	Less than -5	-5 to +5	6-39	40-69	70-79	80-84	85-89	90-94	95-99	100	
70												
65												
60												
55												
50												
45												
40												
35												
30												
25												
20												
15												
10												
5												
0												
Total												

Rate of work satisfactory.

Diagnosis : Rate of reading too slow
Additional training needed.

Median Number of Last Question Answered.....

Median Index of Comprehension.....

Rate of work satisfactory.

Diagnosis: Rate of reading too slow
Additional training needed.FIG. 5. SHOWING FORMS USED IN RECORDING THE SCORES OBTAINED BY
USING COURTIS'S SILENT READING TEST NO. 2

having a score of 5 to 9 questions answered, and similarly for the other intervals. Then each of these piles is sorted according to the "index of comprehension." Suppose there are seven cards in the pile of forty-five to forty-nine questions answered and that these have the following "indices of comprehension": -15, 20, 47, 58, 73, 79, 80. The entries for these scores would be made on the "45 line." To record the -15 score, a 1 is placed in the "Less than -5" column; to record the 20 score, a 1 is placed in the "6-39" column; to record the 47 and 58 scores, a 2 is placed in the "40-69" column; and so on. Courtis emphasizes the necessity of using this record sheet in interpreting the "number of questions answered" and the "index of comprehension." When these two scores are treated as independent measures, the interpretation is likely to be misleading if not erroneous.

Function. The function of the Courtis Silent Reading Test No. 2 is defined as being "to measure ability to read a simple story and to comprehend its simple elements and their simple relationships." In another place the author states that it is intended to measure "the ability to read silently and understand a simple story and simple questions about the story." In correspondence with the writer, Courtis has stated "I know the test is not a valid measure of reading ability except in grades two, three, and four; that at the upper levels of reading ability the scores reflect rate of motor activity much more than reading ability." The fulfillment of its announced function is dependent upon close conformity with the directions prepared by the author. This is an important consideration in all tests but is especially important in this case.

Limitations. The Courtis Silent Reading Test No. 2 presents the pupil with a more normal reading situation than is found in the other silent reading tests which are described in this book. However, only the rate of reading is measured in

Part I, in which the continuous story appears. The comprehension is measured in the second part and the pupil has the opportunity to re-read the story. Since a pupil's rate score and his comprehension scores (number of questions answered and index of comprehension) are derived from separate performances, we cannot know, except indirectly, concerning his comprehension in the first part of the test. It is possible that the pupil failed to understand much of the story in the first reading. It is intended that the number of questions answered and the index of comprehension will furnish an indication of his comprehension in Part I.

The function of the test is narrow. It is restricted to one type of silent reading. Above the fourth and fifth grades the pupil meets many demands for other types of material, especially in the "content subjects." Hence this test should not be used above the sixth grade and will generally be of little value above the fourth grade.

Norms. The author has given the following norms. No statement is made concerning the time of year for which they were calculated:

GRADE	2	3	4	5	6
Rate.....	84	113	145	168	191
Questions.....	16	24	30	37	40
Index of comprehension.....	59	78	89	93	95

4. Burgess Picture Supplement Scale

General structure. The Burgess Picture Supplement Scale is designed for Grades III to VIII.¹ There are four forms. It consists of twenty exercises in which the pupil

¹ It has been used in the second grade.



1. Here is a picture of a girl's head. Take your pencil and quickly draw a circle around the picture, to make a frame for it. Do not spend time trying to make a very good circle; but draw it quickly the first time; and then go on and read what the next paragraph tells you to do.



2. This sleepy woman has not yet finished dressing. Take your pencil and blacken one of her feet; so that it will look as if she had put on one of her shoes. Do not blacken the other foot; because, if you do, she will grow lazy, and will expect some one to help put shoes on her feet every morning.



3. Now take your lead pencil and make a line to represent a rope fastened to this big balloon. Draw it down through the print; and make one end fastened to the bottom of the balloon, and the other end fastened to the black line which separates this paragraph from the next one which is just below it.

reads directions for making certain supplementary drawings in connection with a picture. The pupil is allowed five minutes to do as many of these exercises as he can. The drawing which he makes is taken as evidence of his comprehension of the paragraph. A pupil's score is the number of exercises done correctly. The samples given on page 113 will illustrate the nature of the test.

In explaining the test the examiner is directed as follows: "Tell children they are to have a test in reading. Hold scale up and explain that each paragraph tells them to do something to the picture above it with their pencils. They must read carefully, to make sure just what they are to do. They are to read and mark the paragraphs in order, starting at the top and working down, through the first, second, third, and so on. They must do as many as they can in five minutes. Make sure that the pupils understand; then tell them to turn papers over and begin. Allow exactly five minutes. Collect papers."

In scoring the test papers the examiner is directed to count each paragraph correct where the marking, no matter how crude, follows instructions. "Count it wrong if the marking does not follow instructions. Remember this is a test of reading, not of drawing."

Provision is made for translating the "point score," number of exercises done correctly, into a "derived score" on a scale of 100 units. A different zero point is used for each grade. The credit to be assigned in each grade for a given number of paragraphs marked correctly is shown in Table XIII. For example, a third-grade child having ten paragraphs right is given a derived score of 80; a fourth-grade child having seven right, 50; and so on. The derived scores in Table XIII are for February 1. To adjust for other periods of the school year, one is directed to add or subtract from each child's mark as follows:

Grade 3	Oct. 1, +6	Dec. 1, +3	April 1, -3	June 1, -6
Grades 4-8	+2	+1	-1	-2

TABLE XIII. CREDIT CORRESPONDING TO EACH NUMBER OF PARAGRAPHS MARKED CORRECTLY IN EACH GRADE, BURGESS PICTURE SUPPLEMENT SCALE

GRADE	NUMBER OF PARAGRAPHS READ AND MARKED CORRECTLY																				GRADE	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19		20
3	0	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..	..	..	..	..	..	3
4	0	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..	..	..	..	4
5	0	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..	..	..	5
6	0	2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..	..	6
7	..	0	2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	..	7
8	..	..	0	2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98	100	8

Norms. The zero points for the derived scores have been chosen in the different grades so that the norm for all grades is 50. However, the pupils within a grade differ widely in

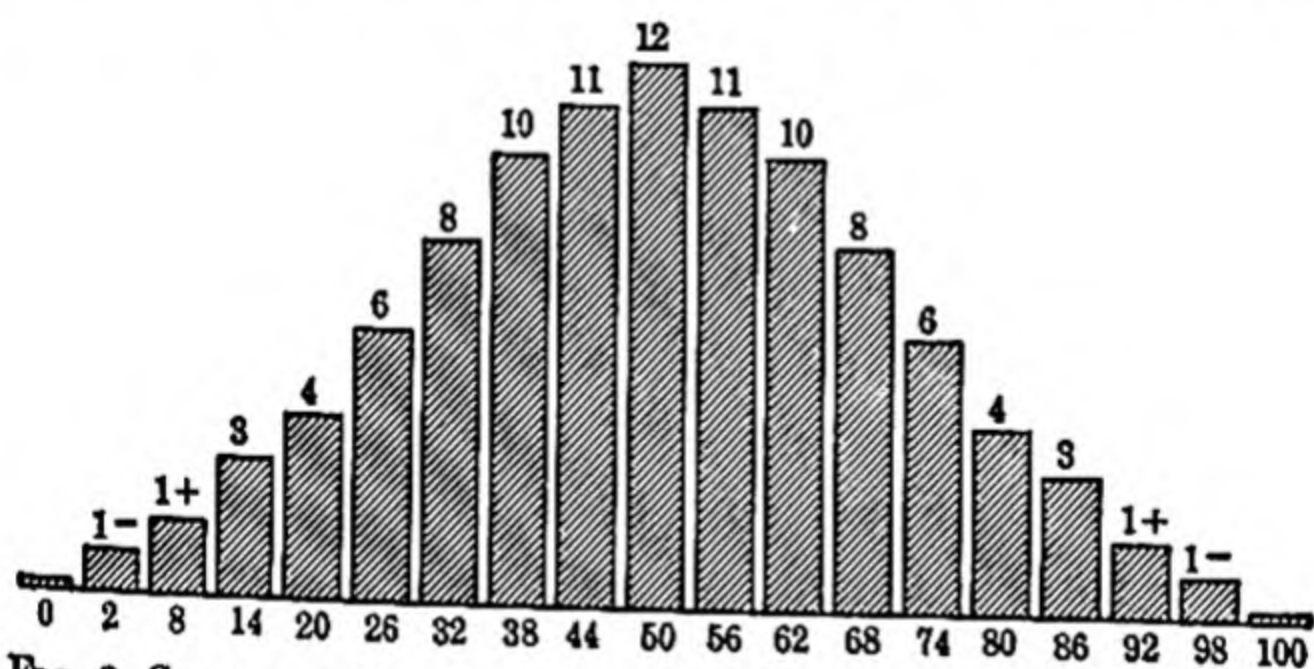


FIG. 6. STANDARD DISTRIBUTION OF SCORES FOR BURGESS PICTURE SUPPLEMENT SCALE

ability. In Fig. 6 the per cent of pupils in the "average grade" receiving each derived score is represented graphically. Figures below columns show marks or credits, and

figures above columns show per cents of children commonly receiving those credits. Thus, twelve per cent usually receive a mark of 50,¹ eight per cent one of 68, and so on. The lowest third in the average class receive marks of from 0 to 38; the middle third, from 44 to 56; and the best third, from 62 to 100.

Limitations. In the derivation of this scale the author identified twenty-five factors which influence a pupil's performance in silent reading. An attempt was made either to eliminate or to control twenty-four of these factors, leaving only the amount read as the variable factor. The ability to draw the pictures required (difficulty of action demanded) is announced as being held constant, but one naturally questions its constancy. Some of the drawings to be made are very simple. For example, in Exercise 6, Form 1, a pupil is asked to cross out a portion of a picture. In other exercises the drawing required is more complex. In Exercise 4 the pupil is to draw a picture of three feathers. No instructions are given to the pupil concerning the quality of the drawings required. One would, therefore, expect to find that some pupils would attempt careful drawings, while others would make very hasty sketches.

The test is printed as a single large sheet which is less convenient than a booklet. The instructions for administering the test are meager, and do not specify the exact explanations to be given to the pupils. The pupils are given no preliminary exercises to acquaint them with the nature of the test. One would, therefore, expect the administration of it to lack objectivity. One is also inclined to question the objectivity of the scoring, even though the directions state that any sort of a drawing which shows that a pupil has followed instructions is to be accepted as correct. In fact some

¹ The interval is from 47 to 53. In the other intervals only mid-points are given.

users of the test have found it necessary to prepare detailed directions for scoring the exercises. Investigation has revealed that the duplicate forms are not equivalent. Unless an appropriate correction is made for this, a constant error will be introduced in the scores yielded by certain forms.

The pupil's score is the number of exercises done correctly in the five minutes allowed for the test. The author considers that both the difficulty of the exercises and the quality of reading required are kept constant. The uniformity of difficulty is secured by use of exercises for which the same per cent of correct responses is in general obtained. The quality of the reading is considered to be kept constant because the pupil is given credit for only those exercises which he answers correctly. Hence, according to the author the score is a measure of the amount of reading which the pupil has done in five minutes or, in other words, his rate of reading.

It is likely that the number of exercises done correctly furnishes the best single numerical description of a pupil's performance on this test. However, this score is not a measure of his rate of work. It is a combination of rate of reading and quality of reading. This combination is not the same for all pupils. For example, a pupil may make a score of 10 doing ten exercises with 100 per cent accuracy. Another pupil may make a score of 10 by doing twenty exercises with 50 per cent accuracy. The rate of reading is not the same for these two pupils. Neither have they read with the same quality.

*5-9. Silent Reading "Scales" requiring answers to questions with text at hand*¹

Facts of title. 5. Scale Alpha for Measuring Understanding of Sentences, devised by E. L. Thorndike, 1914. This

¹ Under this head five silent reading tests are described together. All of them are "scales" and are similar in structure. There is no attempt to secure a measure of the rate of reading. See Chapter X for a description of the silent reading "scales" included in the Stanford Achievement Test.

test is for use in Grades III to VIII. There is no duplicate form.

6. Scale Alpha 2 for Measuring Understanding of Sentences, Parts I and II, devised by E. L. Thorndike, 1915. This is a revision of Scale Alpha and is intended for use in the high school as well as in the elementary grades.

7. Achievement Examination in Reading, Sigma 1, devised by M. E. Haggerty, 1920. This scale is for use in the first three grades. It was used in the Virginia State Survey.

8. Achievement Examination in Reading, Sigma 3, by M. E. Haggerty, 1920. This scale is designed for use in Grades V to XII. It consists of three parts. A vocabulary test, a sentence test, and a paragraph test.

9. Thorndike-McCall Reading Scale by E. L. Thorndike and W. A. McCall, 1920. This scale is designed for use in Grades III to VIII. There are ten forms.

Nature of pupil's performance. In these tests the pupil answers questions or does other exercises with the text at hand so that he may refer to it as frequently as he desires. The material to be read consists of isolated paragraphs. A paragraph with the questions based on it forms an exercise. The exercises are arranged in ascending order of difficulty to form a scale. The Haggerty Achievement Examination in Reading, Sigma 1, consists of two parts. The first part is made up of exercises which are directions for adding certain points or lines to drawings given in the test. The test is rather similar to a picture completion test. The second part is essentially an information test. The questions are to be answered by either "yes" or "no." Part II of the Haggerty Achievement Examination in Reading, Sigma 3, which bears the sub-title "Sentence Reading," is also an information test of the same type.

The time allowance for both Sigma 1 and Sigma 3 is

twenty minutes for the portion of the test devoted to paragraph reading; for Thorndike-McCall Reading Scale, thirty minutes, and for Thorndike Scale Alpha 2, forty minutes. These time allowances are sufficient for practically all pupils to complete all of the exercises which they have the ability to do. In fact many pupils will finish these tests before the end of the period allowed. Hence no measure of rate of reading is secured.

Thorndike tells us that in the construction of the exercises of his Scale Alpha "a prime consideration has been that the paragraph (and questions) be fairly representative of reading in general, even at the cost of some inconvenience in scoring." The determination of what paragraphs and questions were "representative of reading in general" appears to have been made without reference to any objective criterion. The nature of the exercises of the scale may be illustrated by Set III, Difficulty 6.

Read this and then write the answers. Read it again if you need to.

It may seem at first thought that every boy and girl who goes to school ought to do all the work that the teacher wishes done. But sometimes other duties prevent even the best boy or girl from doing so. If a boy's or girl's father died and he had to work afternoons and evenings to earn money to help his mother, such might be the case. A good girl might let her lessons go undone in order to help her mother by taking care of the baby.

1. What are some conditions that might make even the best boy leave school work unfinished?
2. What might a boy do in the evenings to help his family?
3. How could a girl be of use to her mother?
4. Look at these words: idle, tribe, inch, it, ice, ivy, tide, true, tip, top, tit, tat, toe.

Cross out every one of them that has an i and has not any t (T) in it.

Read this and then write the answers to 5, 6, and 7. Read it again if you need to.

Nearly fifteen thousand of the city's workers joined in the parade on September seventh, and passed before two hundred thousand cheering spectators. There were workers of both sexes in the parade, though the men far outnumbered the women.

5. What is said about the number of persons who marched in the parade?.....
6. What did the people who looked at the parade do when it passed by?.....
7. How many people saw the parade?.....

Thorndike Scale Alpha is the model after which the other tests of this type have been constructed, and, with the exception of Haggerty's Sigma 3, there has been practically no modification in the general character of the exercises. The following exercises are quoted from the Thorndike-McCall Reading Scale, which is the most recently published scale of this type.

Read this and then write the answers. Read it again if you need to.

According to the *Kansas City Star*, the wheat farmers of Kansas are too prosperous to trouble themselves about careful harvesting. They do not cut the fields clean. A gleaner eighty years old, after the wheat harvest in Pawnee County last year, went over the wheat-fields with a wagon, a rake, a brush, and a shovel and swept up the wheat left on the ground by the threshers. He gathered nine hundred bushels in forty days, and sold it at a dollar a bushel.

23. Might a farmer be prosperous and still have wheat swept up after the threshers?.....
24. Is this story about the farmers of Arkansas?.....
25. Might a farmer be prosperous and still waste a hundred bushels of wheat?.....

Read this and then write the answers. Read it again if you need to.

There are two methods by which one might make himself acquainted with anything made up of related parts; as, for example, a watch. He might take the watch apart, piece by piece, and while doing so study the details of its structure and the relation of

its parts one to another. An operation like this, which begins with the whole and descends to the parts which compose the whole, is called analysis. The word means a taking apart or separating. Or he might begin with the parts, and, after some experiment and study, get an excellent knowledge of the watch by putting its parts properly together. An operation of this kind is called synthesis.

29. Name in order the method which (a) is easiest, (b) requires most originality.
30. Experimentation is more essential with which process?
31. Copy the words which tell what a mechanism is.

Haggerty Reading Examination, Sigma 3, is "the outgrowth of studies in reading materials and examinations extending over a number of years and of methods developed in the survey of public schools of St. Paul and in the State survey of North Carolina. . . . The material for both Sigma 1 and Sigma 3 examinations has been selected after careful study of the contents of school readers and, in the case of Sigma 3, of textbooks in United States history intended for the seventh and eighth grades. With the exception of the final paragraph of Sigma 3, neither of the examinations contains anything not warranted by general usage in the large number of books examined." The exercises also differ from those of the other tests with respect to the performance required of the pupil. Instead of writing answers to questions, he checks statements which are "true" or "false." A statement is considered "true" if it is in agreement with the paragraph. A "false" statement is one which does not agree with the paragraph read. Exercises I and VI are reproduced to illustrate this test.

A carriage, drawn by four horses, dashed 'round the turn of the road. Within it, thrust partly out of the window, appeared the face of a little old man, with a skin as yellow as gold. He had a low forehead, small, sharp eyes puckered about with innumerable wrinkles, and very thin lips, which he made still thinner by pressing them forcibly together.

- b. A knight's page could fight.
- c. A vanquished knight retained his horse.
- 4. Check the true statements:
 - a. Champions were prohibited to use the sword.
 - b. An unhorsed knight could renew the fight.
 - c. An opponent was vanquished if his arms touched the palisade.
 - d. A knight dragged from the lists by his page was beaten.

In Sigma 1, and Sigma 3, Haggerty has provided a preliminary exercise to acquaint the pupils with the nature of the test. The directions to the pupils in the Thorndike-McCall Reading Scale are put in the form of an exercise. In this way it fulfills the double purpose of acquainting the pupils with the nature of the test and instructing them as to their methods of work. In Thorndike's scales there is no preliminary exercise to acquaint the pupils with the nature of the test,

Description of pupil's performance. These tests consist of exercises arranged in order of increasing difficulty. Thus the logical method of describing a pupil's performance is in terms of the degree of difficulty of the highest point on the scale, which he does with a fixed standard of accuracy (comprehension). This method was followed by Thorndike in Alpha and Alpha 2. It, however, involves a rather intricate statistical procedure and is difficult to use even when the correction tables designed to accompany the scales are at hand. For this reason many users of Thorndike's Scale Alpha have described the performances of pupils in terms of the number of questions answered correctly. Haggerty has followed this method in both Sigma 1 and Sigma 3. It is also the plan of scoring for the Thorndike-McCall Reading Scale. The questions obviously vary widely in difficulty, but this method gives a pupil as much credit for answering correctly one question as for answering correctly any other. However, it is doubtful if a plan of weighting the exercises on

the basis of their difficulty would materially increase the accuracy of the scores.

Derived scores for the Thorndike-McCall Reading Scale. In the Thorndike-McCall Reading Scale, provision is made for translating the point score (number of questions answered correctly) into a T-score. The unit of this derived score called T, in honor of Thorndike and Terman, is one tenth of the standard deviation ($.1 \sigma$) of the distribution of the silent reading ability (as defined by this scale) of all twelve-year-old pupils taken without regard to their grade placement. This distribution is assumed to be included between -5.0σ and $+5.0 \sigma$. Hence a scale of 100 units is provided with the zero point at 5.0σ below the average silent reading ability of twelve-year-old pupils. In order to describe a pupil's performance in terms of this scale, it is first described in terms of the number of questions answered correctly. The basis of translating this score into its T-score equivalent is the per cent of twelve-year-old pupils exceeding, plus half of those answering each number of questions correctly. For example, 95.6 per cent of the twelve-year-old pupils answered more than ten questions correctly.¹ The point on the base line of the distribution reading ability of twelve-year-old pupils above which there are 95.6 per cent of the pupils is 33 (3.3σ above the point 5.0σ below the average). A score of 33 is, therefore, given to pupils who answer ten questions correctly. A table has been prepared which gives the score that corresponds to each number of questions answered correctly, and a user of the scale can easily translate the point scores into T-scores without understanding the statistical procedure just described.

It is claimed for this method of describing a pupil's performance that the unit (a measure of the variability of

¹ More accurately, 95.6 per cent includes half of those who answered exactly ten questions correctly.

twelve-year-old pupils) can be universally used, since it is based upon the variability of a fixed group. It is true that the group is fixed, but it does not follow that the distribution of its ability, and hence its variability, is fixed even in the case of silent reading. It is conceivable that improved methods of instruction and a different type of school organization might materially modify the distribution of reading ability of twelve-year-old pupils.

McCall also ascertained the average T-score for the pupils of each chronological age group. These averages were then used as a basis for a second derived score. For example, the average T-score for pupils having a chronological age of 121 months (ten years and one month) is 40. Hence, a pupil having a T-score of 40 is said to have a "reading age" of 121 months. This merely means that he has a silent reading ability (as measured by this scale) equivalent to the average silent reading ability of all pupils whose chronological age is ten years and one month. A "reading quotient" is obtained by dividing the pupil's reading age by his chronological age. Both of these derived scores have certain merits in the interpretation of the measure of silent reading ability.

It should be noted that the Thorndike-McCall method for describing a pupil's performance in terms of a T-score does not explicitly include any scheme of weighting the questions on the basis of their difficulty. Its use, therefore, is not restricted to scaled data. Also duplicate forms of the test are not required to be exactly equivalent to yield comparable scores provided a translation table has been devised for each form.

Function. The function of Thorndike Scale Alpha 2 is defined as the measurement of "achievement in paragraph reading," which in turn is defined as "that thing much of which enables an individual to respond correctly to a paragraph and a question about the paragraph involving much

'difficulty for paragraph reading,' whereas an individual of less 'achievement' should respond correctly only to a 'paragraph and a question about the paragraph of less difficulty.' Apparently this statement of the function was adopted by the authors of the other tests in this group. In contrast with the other silent reading tests described in this chapter this group may be described as "power tests." It should be noted that Thorndike defined ability to read (or to understand paragraphs) in terms of his test, which is a narrow concept of reading ability.

Norms. The published grade norms for the scales by Haggerty and Thorndike-McCall are given in Table XIV.

TABLE XIV. NORMS FOR HAGGERTY SIGMA 1 AND SIGMA 3, AND THORNDIKE-McCALL READING SCALES

SCALE	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
Sigma 1												
Test 1.....	4	12	16	20								
Test 2.....	2	8	14	18								
Sigma 3					31	50	68	76	84	90	96	102
Thorndike-McCall.....		28	35	41	46	52	57	60	62	63	65	67

More detailed norms are given in the manual of directions which accompany the scales. The norms for silent reading scales by Haggerty and the Thorndike-McCall Reading Scale are computed from the norms announced for the end of each half-grade by taking averages. Hence, they may be considered as norms for April testing. Presumably these norms for the Thorndike-McCall Reading Scale are derived from scores yielded by Form 1. Hence, it is necessary to inquire into the equivalence of other forms when using these norms to interpret the scores yielded by them.

Limitations of these "scales." Except for Thorndike Alpha and Alpha 2, the directions to examiners are sufficiently detailed so that the giving of the tests is probably highly objective. Even in the case of Thorndike's two scales it does not appear likely that different examiners would secure markedly different results. Therefore, the source of any lack of objectivity must be sought in the scoring of the performance of the pupils. Except in the Haggerty Achievement Examinations the pupil is required to express his answers to the questions in terms of a word or a group of words. When pupils are called upon to express themselves in this way, such a variety of responses is obtained that it is extremely difficult to determine which ones should be accepted as correct and which ones should be called wrong. Thorndike gives a list of answers which are to be accepted as correct and also a list of answers which are to be counted wrong. But even with this assistance, the scorer is frequently called upon to exercise judgment and in such cases the scoring is not highly objective. In the Thorndike-McCall Reading Scale the questions require relatively simple answers and the directions for scoring appear to be more complete. Hence, the scoring of this scale is probably more objective. The type of exercises which Haggerty has used in both Sigma 1 and Sigma 3 is such that a very high degree of objectivity is attained, particularly in the case of Sigma 3.

Haggerty has computed the reliability for both Sigma 1 and Sigma 3 by having the same test repeated. In the case of Sigma 1 the interval between the two applications of the test was six weeks. For 200 children in Grades I to III a coefficient of reliability of .84 was obtained. In the case of Sigma 3 the interval between the two applications was only two days. For 126 pupils in Grades V C to VIII A the coefficient of reliability was found to be .885. For the sentence test alone the reliability coefficient was .769. For

the paragraph test it was .806. Gates ¹ reports reliability coefficients for the Thorndike-McCall scale ranging from .25 to .72. These are based upon the pupils belonging to a single grade.

In all of these tests the pupil's performance is limited to the answering of questions with the test at hand so that he can refer to it whenever he desires. Therefore, these tests should not be thought of as measuring reading ability in general, but rather measuring that type of reading ability which is called for in answering questions with the text at hand. The rate of work, which is an important dimension of the silent reading ability, is entirely neglected by this group of tests. This means that no information is obtained concerning the silent reading mechanism. Pupils may lack good eye-movement habits or be afflicted with a high degree of vocalization and yet make high scores on these tests. Hence, at best these tests can yield only a partial measure of silent reading ability.

One writer who has experimented with this type of silent reading test reports that "if the children were given sufficient time they were able to follow instructions couched in the most recondite and unusual phraseology. The only way to make the hurdles high enough was to introduce complexities of thought and difficulties of tasks which tended to make the test one for qualities other than ordinary reading ability." ² This investigator abandoned this type of test.

Although it is expected that the pupil will find it necessary to read the paragraph in order to answer the questions, an analysis of these tests shows that this is not always necessary. For example, the first three questions of Set III of Thorn-

¹ Gates, Arthur I., "An Experimental and Statistical Study of Reading and Reading Tests"; in *Journal of Educational Psychology*, vol. 12, p. 379. (October, 1921.)

² Burgess, May Ayres, *Measurement of Silent Reading*, p. 85. Department of Education, Russell Sage Foundation, New York, 1921.

dike Scale Alpha 2 might be answered without any reference to the paragraph (see page 119). In case a pupil does this, he may not answer them correctly, but the fact remains that some pupils are likely to attempt to answer them in this way. If this should be found to happen in a very large number of cases, we should have reason to doubt the validity of the test.

This criticism also applies to the Thorndike-McCall Reading Scale. For example, the answer to Question 23 is "yes" (see page 120). If this question were incorporated in a general information test, it is likely that a large per cent of correct responses would be obtained. The same answer is required for Question 25, and doubtless the same result would be obtained for it. This criticism appears to apply to six out of thirty-five questions in Form 1 of this scale. In the case of Haggerty's Sigma 1 and Sigma 3, a careful reading of the paragraph is a prerequisite for correct answers except as a matter of chance. In both of these tests a specific section is set aside for testing a pupil's general information.

The standards furnished by the grade norms for these tests suggest that an objective of the school is to train the pupils to read as difficult material as possible. This objective is not in accord with either our educational practice or theory. This constitutes a serious limitation of the usefulness of the tests particularly below the sixth grade. In these grades emphasis is placed upon the development of good silent reading habits, especially eye-movements and elimination of vocalization. Above the sixth grade the mechanism of silent reading is important, but a shifting of the major emphasis to other phases of silent reading may be justified if the training in the lower grades has been effective. Thus these tests are most useful above the sixth grade.

II. VOCABULARY SCALES

The problem of measurement. A fundamental factor of one's ability to read silently is the range of words whose meaning he recognizes. For diagnostic purposes it would be very helpful to have an instrument with which to measure the extent of a pupil's vocabulary, but in the construction of a vocabulary test we encounter difficulty in securing a satisfactory performance from the pupil. It is not satisfactory to have him write out a definition for each word because such a performance cannot be objectively scored. In addition, such a test would require a large amount of time on the part of both the pupil and the scorer. To overcome these difficulties several ingenious types of performances have been used by the makers of vocabulary tests. Five representative tests will be described together.

Facts of title. 1. Thorndike Visual Vocabulary Scale A, devised by E. L. Thorndike, 1914. The revised scales, A_{2x}, A_{2y}, B_x, and B_y were published in 1916. They are designed to be used in Grades II to VIII.

2. Southington-Plymouth English Vocabulary Scale, devised by E. C. Witham, 1919. There is only one form, which is designed to be used in Grades III to XII.

3. Holley Sentence Vocabulary Scale, devised by C. E. Holley, 1919. There is only one form which is designed to be used in Grades II to XII.

4. Haggerty Vocabulary Test, devised by M. E. Haggerty, 1920. This scale is one of the sub-tests of the Haggerty Reading Examination, Sigma 3. It is designed for Grades V to XII. There is only one form.

5. Thorndike Test of Word Knowledge, devised by E. L. Thorndike, 1921. There are four forms and the test is designed to be used in Grades II to VIII.

Nature of pupil's performance. In Scale A and Scale A 2,

Thorndike asks the pupil to classify the test words. The pupil is directed:

Look at each word and write the letter F under every word that means a flower.

Then look at each word again and write the letter A under every word that means an animal.

Then look at each word again and write the letter N under every word that means a boy's name.

Then look at each word again and write the letter G under every word that means a game.

Then look at each word again and write the letter B under every word that means a book.

Then look at each word again and write the letter T under every word like now or then that means something to do with time.

Then look at each word again and write the word GOOD under every word that means something good to be or do.

Then look at each word again and write the word BAD under every word that means something bad to be or do.

The Holley Sentence Vocabulary Scale consists of exercises like the following. The pupil is asked to draw a line under the one of the last four words which makes the truest sentence:

A gown is a string animal dress plant
 An orange is a dress animal fruit hornet
 Envelopes are made for letters snakes water apples
 Haste is hurry red little sweet
 Scorch means to cut burn bruise turn

The Haggerty Vocabulary Test consists of exercises that are somewhat similar to those used by Holley. Samples are given below. The pupil is directed to "Draw a line under the word or phrase which is the best definition of the first word."

1. minister (servant, preacher, agent, to assist)
2. student (one who seeks knowledge, teacher, paper, book)
3. pardon (forgive, hinder, condemn, smile at)
4. island (section, part of the ocean, land surrounded by water, peak)
5. float (sail, sink, to fly, to stay on top of the water)

Witham gives fifty words and their definitions on opposite pages. The definitions are not arranged in the same order as the words. The pupils are required to match up the words and definitions.

The Thorndike Test of Word Knowledge is very similar in structure to the Haggerty Vocabulary Test. In all cases the pupil is to underline the word "which means the same or nearly the same." Five possible words are given.

Selection of test words. The words used in these vocabulary tests have been selected in a variety of ways. Holley used a list prepared by Terman and Childs, who took the "last word of every sixth column" in Laird and Lee's Vest-Pocket Webster's Dictionary, 1904 edition. Thorndike selected the words of his first test on the basis of difficulty from a list which was suitable for classification. Presumably the same method was followed in selecting the words for the new test. The words on which the exercises of the Haggerty Vocabulary Test are based were selected in much the same way. Witham selected the words for the Southington-Plymouth English Vocabulary Scale from the list of one thousand words in the Ayres "Measuring Scale for Ability in Spelling" and from the words in "Thorndike's Reading Scale A-2, Word Knowledge or Visual Vocabulary" (X series), and from the Provisional Extension of that scale. The first ninety words are from the Ayres Scale; the last ten are from Thorndike's. The scheme for selecting the words from the Ayres Spelling Scale was as follows:

First	250 — every 50th word chosen	5
Second	250 — every 25th word chosen	10
Third	250 — every 10th word chosen	25
Fourth	250 — every 5th word chosen	50
		90

Beginning with line 8½ X on the Thorndike Scale A-2, one word was chosen from each remaining line and one word from each line

on the Provisional Extension of this scale. This gives ten words, making in all the one hundred words in the test.

Arrangement of test words. In all of these tests the words appear to be arranged in order of increasing difficulty, but, except in the case of those in which the words were selected on the basis of difficulty, definite evidence of the exact procedure is wanting. The Thorndike Scale A has five words for each level of difficulty. Scale A-2 and Scale B have ten words for each step. The time allowance for these scales is such that practically all pupils are able to finish all of the exercises they are able to do correctly.

Description of pupil's performance. In his first test Thorndike directed that the pupil's score be expressed in terms of the degree of difficulty of the list of words in which the pupil classifies 80 per cent correctly. When a pupil fails to have exactly 80 per cent correct of any group of words, tables are provided for inferring his score. In the Nassau County Survey, the pupil's score was taken as the number of words correctly marked. This is the simplest method of computing a pupil's score. In all of the other tests the pupil's score is the number of exercises done correctly. This is the number of words which have been defined in the way required by the test.

Function. As the title implies, the function of vocabulary tests is to *measure acquaintance with the words* of the English language. The tests which have been described differ in the definition of "acquaintance with words." Thorndike states, "The obvious purpose of these scales is to measure how hard words a pupil can read in the sense of understanding their meaning well enough to classify them under the proper heading." Although different definitions of "acquaintance with words" are indicated by the differences in the structure of the tests, it is likely that all of the authors hoped to secure a measure of that acquaintance with words

which is required for understanding sentences and paragraphs. This probable function, however, is indefinite because various degrees of acquaintanceship are required for reading, and the difficulty of the material to be read depends upon the sentence structure as well as upon the vocabulary. Therefore, a precise statement of the function is not possible.

Limitations. These vocabulary tests are highly objective. Only one answer can be accepted as correct. The reliability of vocabulary tests has been given little attention. Haggerty reports a coefficient of reliability of .865 based upon the repetition of his test (not duplicate form), after an interval of two days, to a group of pupils in Grades V to VIII. McCall, using Thorndike's Visual Vocabulary Scale A and a second scale similar to it, reports a coefficient of reliability of .53 based upon 88 pupils in the sixth grade. Wyman and Wendell report a reliability coefficient of $.79 \pm .04$ for Thorndike's Visual Vocabulary Scale B (Series X and Y).

The available evidence furnishes little basis for even hypotheses concerning the validity of the various vocabulary tests as instruments for measuring one's acquaintance with the words of the English language. In all of these tests the performance required of the pupil differs from that occurring in associating meanings with words as they are met in reading. The words are not presented in context, but in isolation.¹ The effect of this is not known. It seems reasonable to expect that pupils will demonstrate acquaintanceship with fewer words when they are presented in isolation than they should if the words were presented in context. Granting this is the case, it may be that the difference approaches a constant. In case it does, the validity of the measures would not be affected for comparative purposes.

¹ It may be urged that Holley presents the words in sentences, but the form of presentation is mechanical and the sentences are not complete. Hence, this test should not be considered an exception to this statement.

The little available evidence with reference to the correlation between vocabulary ability and reading ability is inadequate, but what we have does not indicate that the vocabulary tests fail to measure a pupil's acquaintance with words.¹ Thorndike states with reference to his revised scales that they are "better than any one would probably devise himself with less than five hundred hours of work." He also gives the opinion that the measurement of word knowledge by classification "seems undoubtedly better than a definition test. Measurement of how hard words a pupil knows is beyond question sounder than measurement by an arbitrary score in a 'blanket' test."

III. GRAY ORAL READING TEST²

General structure. The normal performance in oral reading can be observed, although for the more subtle qualities of expression this observation will not be very accurate. For this reason the measurement of ability in oral reading has been limited to the accuracy and fluency of pronunciation. For measuring these elements of the ability of pupils to read orally, Gray has devised a test consisting of a series of paragraphs arranged in order of gradually increasing difficulty in oral reading. The test is designed to be used in all grades beginning with the first. The nature of the test can be best illustrated by reproducing a few of the paragraphs:

1. A boy had a dog.
The dog ran into the woods.

¹ See Gates, Arthur, "An Experimental and Statistical Study of Reading and Reading Tests"; in *Journal of Educational Psychology*, vol. 12, p. 457. (November, 1921.)

² As the proof of this chapter is being read the publication of the Gray Oral Reading Check Tests is announced. These tests differ in certain respects from the test described here. They are designed primarily for diagnosis and no composite score is obtained. The manner of giving the test is however the same as described here.

The boy ran after the dog.

He wanted the dog to go home.

But the dog would not go home.

The little boy said, "I cannot go home without my dog."

Then the boy began to cry.

6. It was one of those wonderful evenings such as are found only in this magnificent region. The sun had sunk behind the mountains, but it was still light. The pretty twilight glow embraced a third of the sky, and against its brilliancy stood the dull white masses of the mountains in evident contrast.

11. The hypotheses concerning physical phenomena formulated by the early philosophers proved to be inconsistent and in general not universally applicable. Before relatively accurate principles could be established, physicists, mathematicians, and statisticians had to combine forces and work arduously.

Giving the test. In giving the test the pupils are taken one at a time and asked to read beginning with the first paragraph. As the pupil reads, the teacher records, on another copy of the test, two sets of facts: the number of seconds the pupil takes to read each paragraph and the errors which are made. To obtain the number of seconds the teacher must have a watch with a second-hand or, better, a stop-watch. Six types of errors are recorded: (1) complete mispronunciation of a word so as to indicate that the pupil has no control over it; (2) partial mispronunciation; (3) omissions; (4) substitutions; (5) insertions; and (6) repetitions. The method of marking these errors on the test paper is illustrated in the following quotation from the class record sheet:

The sun pierced into ^{many} my large windows. It was the opening of October, and the ^{clear} sky was of a dāzzling blue. I looked out of my window and down the street. The white house of the long, straight street were amost painful to the eyes. The clear atmosphere allōwed full play to the sun's brightness.

If a word is wholly mispronounced, underline it as in the case of "atmosphere." If a portion of a word is mispronounced, mark appropriately as indicated above: "pierced" pronounced in two syllables, sounding long *a* in "dazzling," omitting the *s* in "houses," or the *al* from "almost," or the *r* in "straight." Omitted words are marked, as in the case of "of" and "and"; substitutions, as in the case of "many" for "my"; insertions, as in the case of "clear," and repetition, as in the case of "to the sun's." Two or more words should be repeated to count as a repetition.

To give the test satisfactorily requires practice in detecting the errors and in recording them. The teacher should have some one read the test, intentionally making errors, so that he may become skillful in detecting each of the six types of errors.

Describing the pupil's performance. Depending upon the time required and the number of errors, the pupil may receive full credit, three fourths credit, one half credit, one fourth credit, or no credit for the reading of each paragraph. Table XV gives in schematic form the amount of credit that is to be given for various performances:

TABLE XV. CREDITS TO BE GIVEN FOR COMPLETE AND PARTIAL SUCCESS IN READING A PARAGRAPH OF GRAY ORAL READING TEST

SECONDS	ERRORS							
	0	1	2	3	4	5	6	7 or more
40 or more.....	4	4	3	2	1	0	0	0
30-39.....	4	4	3	2	1	1	1	0
25-29.....	4	4	3	2	2	1	1	0
20-24.....	4	4	3	3	2	1	1	0
19 or less.....	4	4	4	3	2	1	1	0

The numbers in the left-hand column refer to the time required to read a paragraph. The numbers, 0, 1, 2, etc., in

the horizontal line at the top of the table refer to the number of errors made in the reading. The numbers in the horizontal line to the right of "40 or more" mean that, if a paragraph is read in 40 or more seconds with no errors, a credit of 4 (full credit) is given; with one error, a credit of 4; with two errors, a credit of 3 (three fourths credit); with three errors, a credit of 2; and so forth. Thus, the credit to be allowed for the reading of a paragraph can be found from the number of seconds required to read it and the number of errors made. For example, paragraph four is read by a pupil in 34 seconds with 3 errors. In the left-hand column of Table XV find the interval containing 34 seconds. Evidently, it is the interval, 30-39. Follow the horizontal line of numbers to the right of 30-39 to the column which represents 3 errors. The credit in this column is 2. This means that the pupil is to receive two fourths credit for this paragraph.

The instructions for finding the total score of a pupil direct the examiner to construct a table of the form shown in Table XVI. In the first column of the table the numbers corresponding to the paragraphs of the test are entered. The credit which the pupil makes on each paragraph is to be entered in the second column. The values of the paragraphs, with the exception of paragraph one, are shown in the third column. Paragraph one has a different value for each grade. These are given in the right-hand margin of the table. The products of the credits times the corresponding values are entered in the fourth column. The sum of these products is divided by 4. The resulting quotient is the pupil's score on Gray's Oral Reading Test.

It is obvious that the scores for pupils in different grades are expressed in terms of scales which differ in respect to the zero point. In the first grade, paragraph one has a value of 55 above the zero point. In Grade two, it has a value of 35, which means that the zero point has been moved up twenty

TABLE XVI. FORM TO BE USED IN CALCULATING INDIVIDUAL SCORES FOR GRAY ORAL READING TEST

PARA- GRAPH	CREDIT	VALUE	PRODUCT OF CREDIT AND VALUE	VALUE FOR PAR- AGRAPH 1
1				Grade I 55
2		5		II 35
3		5		III 30
4		5		IV 25
5		5		V 20
6		5		VI 15
7		5		VII 10
8		5		VIII 5
9		5		
10		5		
11		10		
12		5		

Total product.....

Pupil's score (Total product divided by 4).....

units. The shift for the successive grades is five units. The use of different scales in the different grades is confusing and makes comparisons between grades difficult. In addition, the construction of a form, such as is shown in Table XVI, for calculating the score of each pupil makes the scoring tedious.

A proposed modification of Gray's method. If the same zero point were used for all grades, it would then be possible to assign fixed values to the paragraphs and to simplify greatly the scoring. For example, if the zero point for the first grade is used for all grades, the several paragraphs then have values as follow:

PARAGRAPH	SCALE VALUE
1	55
2	60
3	65
4	70
5	75
6	80
7	85
8	90
9	95
10	100
11	110
12	115

With these scale values assigned to the paragraphs, a pupil's score in any grade may be defined as follows: a pupil reading satisfactorily a number of successive paragraphs and failing of complete success on all more difficult paragraphs receives as his score the scale value of the most difficult paragraph read successfully, as defined in Table XV, plus additional credits for partial success on more difficult paragraphs; a pupil failing of complete success on a paragraph and then reading a more difficult one satisfactorily receives as his score the scale value of the most difficult paragraph below which there are no failures, plus the additional credits for partial and complete successes. The computation of a pupil's score by this method is illustrated by the following case. Suppose that a pupil has read successfully all paragraphs up to and including the fifth. On paragraph six he is to receive 2 credits and on paragraph seven, 1. Paragraph five has a value of 75. Thus, the score would be 75 plus two fourths of 5 plus one fourth of 5, or $78\frac{3}{4}$.

The calculation of a pupil's score is unnecessarily difficult because of the presence of fractions. These can be eliminated by the simple device of using a larger unit. If we increase the unit of the scale so that the interval between suc-

cessive paragraphs is 4 instead of 5, Table XV will give exact corrections to be added. Therefore, it is proposed that, instead of the scale values as given above, the following be used:

PARAGRAPH	NEW SCALE VALUE
1	44
2	48
3	52
4	56
5	60
6	64
7	68
8	72
9	76
10	80
11	88
12	92

The finding of a pupil's score in terms of this new scale may be illustrated by the following record:

PARAGRAPH	VALUE	SECONDS	ERRORS
1	44	22	0
2	48	27	0
3	52	28	1
4	56	33	1
5	60	37	2
6	64	35	3
7	68	40	5
8	72	40	7
9	76	45	7
10	80		
11	88		
12	92		

According to Table XV, the most difficult paragraph which this pupil read successfully is the fourth. The same table gives 3 credits and 2 credits for paragraphs five and six. The pupil's performance on more difficult paragraphs is considered to be a complete failure. Therefore, the pupil's score is 56 plus 3 plus 2, which equals 61. In addition to the simplicity of its calculation, this score has a very definite interpretation in terms of the test. It means that the pupil's oral reading ability is slightly more than that represented by the successful reading of the fifth paragraph, which has a scale value of 60. It is true that the pupil did not read this paragraph with complete success, but he read the sixth paragraph with partial success which accounts for the increase in his score.

The method proposed for computing the individual score is logically consistent with that given by Gray, except for a pupil who fails to read successfully the first paragraph. In fact, most of the procedure is implied in the unique method of graphical representation which he has used. In this representation he essentially corrects for the shift in the zero point by making a corresponding shift in the horizontal axis of the diagram for the successive grades. From one point of view this is equivalent to using the same zero point for all grades. In the case of a pupil who fails to read the first paragraph successfully, the credit assigned by Gray's plan depends upon the pupil's grade placement. He receives a fractional part of the credit given for the first paragraph. The proposed plan does not provide at all for the calculation of scores of pupils who fail to read the first paragraph successfully. To be consistent with Gray's procedure, the pupil who earns 3 credits on the first paragraph would receive as his score three fourths of 44 plus additional credits for partial successes on more difficult paragraphs. This plan, however, appears to penalize unduly pupils who fail to read the first paragraph successfully.

Norms. Grade norms for both methods of scoring are given in Table XVII.

TABLE XVII. NORMS FOR GRAY'S ORAL READING TEST

GRADE	REVISED SCORES	GRAY'S METHOD
1.....	25	31
2.....	50	43
3.....	57	46
4.....	62	47
5.....	66	48
6.....	71	49
7.....	74	47
8.....	78	48

Limitations of Gray's Oral Reading Test. No information is at hand concerning the reliability of the Gray Oral Reading Test. Since there is only one form available, independent duplicate measurements cannot be made. It appears likely that the test approximates in reliability the silent reading tests described in this chapter.

An obvious practical limitation is the necessity of administering the test to one pupil at a time. For this reason the time expenditure for testing an entire class is prohibitive except for very enthusiastic workers.

IV. INTERPRETING SCORES AND PLANNING REMEDIAL INSTRUCTION

Interpretation of scores of individual pupils. Space does not permit a consideration of the interpretation of scores of individual pupils. This topic has been treated in the author's book *Measuring the Results of Teaching* and more extensively in a recent monograph by Professor W. S. Gray¹

¹ Gray, William Scott, *Remedial Cases in Reading: Their Diagnosis and Treatment*. Supplementary Educational Monograph no. 22. (Chicago: University of Chicago, 1923. 208 pp.)

and a book by Professor C. T. Gray.¹ The analytical and supplementary diagnosis described and illustrated in these references will be found helpful in dealing with children who have difficulties in learning to read.

Limitations which should be kept in mind. Prior to the derivation of standardized silent reading tests this field of achievement was almost entirely neglected; consequently, silent reading tests render a valuable service if they do nothing more than direct attention to this phase of school work. The fact that all phases of silent reading ability are not measured and that the measures are subject to both constant and variable errors is not sufficient reason for rejecting the standardized silent reading tests which are now available. However, the measures of achievement yielded by reading tests must be used intelligently if the best results are secured. The limitations which have been noted in the preceding pages must be kept in mind.

Interpretation of scores of classes. Space does not permit a detailed consideration of the interpretation of the scores of a class, but the general plan to be followed is similar to that described for arithmetic beginning on page 76. When rate and comprehension are measured separately, five cases will occur. When only comprehension is measured or the two dimensions are measured together, there will be only three cases: (1) class score below standard; (2) scores scattered too wide; (3) class score up to or above standard. In all cases the interpretation must be made in terms of the instructional needs of the pupils. Remedial instruction must then be adapted to these needs. Space does not permit a consideration of the remedial instruction suitable to certain needs. In addition to the three sources referred to above in connection with remedial instruction for individual

¹ Gray, Clarence Truman, *Deficiencies in Reading Ability: Their Diagnosis and Remedies*. (New York: D. C. Heath and Company, 1922. 420 pp.)

pupils, a recent book by Stone will be found helpful.¹ For purposes of illustration the remedial instruction for the first case mentioned above will be considered.

To raise the median comprehension score of the class. Suppose that a silent reading test has revealed an unsatisfactory median comprehension score, but a satisfactorily close grouping of the scores around the median. The commonest of all reasons for this situation, particularly when found in the grades below the sixth, is that the teachers have been placing chief stress upon oral reading. Where children are required to give their attention mainly to the correct pronunciation of words, the correct enunciation of sounds, and the correct inflection of the voice in passing over the several punctuation marks, not much growth in the power to comprehend meaning in the language can be expected. Where the children study their reading lesson with the point of view of being able to respond in this way, they fasten upon themselves the habit of watching for words whose pronunciation they are not sure of, or they form the habit of reproducing the sounds of syllables, thus establishing the practice of moving the lips and other speech organs when reading silently. Frequently both these habits fix themselves upon children whose reading is judged mainly by the daily oral performance. When either or both habits become fixed a real struggle is required to break them. Unless they are broken, however, the child suffers a severe handicap the rest of his reading life. Many men and women of mature years are still paying the price of those habits fixed in youth. They read but little faster silently than they can pronounce the words orally, because their speech organs make all the motions of the successive words as the reading proceeds.

Care from the beginning. To be on guard against these

¹ Stone, C. R., *Silent and Oral Reading*. (Boston: Houghton Mifflin Company, 1922. 306 pp.)

two habits care must be exercised from the very beginning. Children in the primary grades should have exercises from the start in which the meaning is the only significant element, and the response is not in terms of words said, but things done, or interpretations made. For example, let it be the usual thing for the child to carry out the directions contained in the word or sentence. The primary teacher should be supplied with some hundreds of cards upon which such sentences or short paragraphs as the following are printed or written:

- (1) Draw a picture of a flag on the blackboard.
- (2) Make a sound like a cross kitty makes when a dog chases her.
- (3) Hide behind the door.
- (4) Play that you are carrying a cup full of water and do not wish to spill any of it.

These cards should be graded in such a way that certain ones will contain only the words taught in the first reading lessons. As more words are learned, more cards will become available. Word drills should then divide time very generously with these practice cards which emphasize attention to meaning.

Variety in handling the exercises may be introduced in scores of ways which will readily occur to a resourceful primary teacher. Many other devices having the same aim will also occur to the teacher. The essential thing is that practice in translating written or printed language into action instead of words should be started early, thus producing the habit of advancing through a paragraph by thought-units rather than by letters, syllables, or words.

Reading above the primary grades. In grades above the primary the problem is fundamentally the same as stated for the primary, but the devices must vary.

First, whenever reading is done orally, be sure that what the child is reading is new to most of his listeners. Be sure,

too, that the other pupils are listening, and not following along with the reader in another copy of the same book. No method of reading is more faulty in intermediate grades than that in which other members of the class are watching for a word error of the reader, ready to call attention at once to such a mechanical mistake. This method centers the attention of the reader constantly upon the mechanics and never develops the habit of attending first to the thought. Whereas, if the reader realizes that his hearers know nothing of the content of his selection except what they gather from his reading, then giving the thought instead of pronouncing the words becomes the controlling factor in his consciousness. It follows from this that only selections, the thoughts in which are vital to children, should be used as subject-matter for such reading. Then let the one who has read such a selection defend the selection against questions or criticisms of the class. In short, center attention upon the meaning, even at the expense, if necessary, of accuracy in pronunciation, enunciation, and expression.

Second, let the amount of reading which is compellingly interesting be increased. Supplementary reading in geography, history, science, and literature should be given a larger place. Require that the reports made upon such readings be rather exact, but let the selections be reasonably easy for the children. Gain in facility in silent reading cannot be secured by holding the children to selections which are so difficult that word-troubles absorb all the attention. One must be able to go with ease through the successive thoughts before the habit of attending to the thought can be acquired.

Third, make all the industrial and playground exercises give a far greater measure of service in teaching reading than they now commonly give. How singularly shortsighted we are to ask a child to follow the directions printed in his arith-

metic for finding the per cent that one number is of another, but employ a teacher to give orally the directions for playing a new game, making a raffia basket, or planting beans. The very things which come nearest the natural interests of the children, concerning which they would most zealously read if they had the paragraphs containing the needed directions, are given to them orally. When interesting school exercises require a careful following of directions, then those directions make the most effective silent reading material. But in practice we seldom make use of them. This fault is due to a failure to understand the distinction between the aim of the intermediate grades and the aim of the upper grades. If we realized that all the work of the intermediate grades should be made to develop skill in using the tools of learning, then we should not conduct these exercises without making them aid in teaching reading.

Reading in the upper grades. Passing now to the situation presented when the score of a class above intermediate grades is found to be low, we have the most serious task of all. The junior high-school or upper-grade pupil should be able to proceed with his school tasks without much attention to the tools he is using. It is not the primary function of this department of the school system to increase the children's facility in the handling of these tools. However, success in nearly all the tasks undertaken in the upper grades depends upon the skill which the children are expected to possess in the tool subjects. A compromise is, therefore, necessary, if children in the junior high school, or seventh and eighth grades, are found deficient in their ability to read silently. A few suggestions are here offered in the hope that some help may come from them, although it is realized that correcting reading faults at this stage is very difficult.

First of all, the children's own conscious efforts should be obtained in the direction of correcting the faults. Then, too,

the teacher should see that he is observing the same fundamental principles stated for the intermediate grades. Comprehension, and not mechanics, must be made the test of all reading, whether in history, science, or literature. The material selected for use must be sufficiently easy so that the children are not tied up in word or language difficulties. Again, to overcome the habit of proceeding by too small units, practice must be afforded in advancing by short sentences or phrases.

In case the trouble seems to be that the children read fluently enough orally, but get little of the thought, introduce a great deal of the sort of reading requiring close attention to the thought. For example, use rule books for football, basketball, and the like for those interested in games; catalog descriptions; directions for making certain stitches; the more involved arithmetic problems, and so on. These things possess a minimum of word difficulty and a maximum of thought difficulty. They require the imagination to construct a picture little by little and hold it up for constant modification as the reading proceeds. Thus, attention is focused on thought.

Where the class appears to have the right habits of reading silently but have had insufficient practice, the obvious suggestion is to give them all the practice possible. Much supplementary reading upon which they make only meager reports, if any, will help. Try to secure as much general home reading as possible. See that an abundance of interesting things is available for reading and stimulate interest by having the children's criticisms of them given before the class.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What are the chief methods by which adults add new words to their vocabularies? Are more new words learned from the context in which they appear, or from the dictionary? What can you say concerning the best way to increase the vocabulary of children?

2. What are some of the other factors besides vocabulary involved in silent reading? In what grades is vocabulary the most important factor? Make some suggestions for guaranteeing the intimate association of the mental concept which a word symbolizes, and the word itself when it is encountered in word drills.
3. What is the significance of rate in reading? Is there any truth in the rather common belief that one who reads slowly "gets more out of what he reads"? If you do not know the answer, can you devise some way to test it out in your class? Compare your own silent reading rate with that of some equally well-educated friends.
4. What are the chief dangers involved in having much oral reading in the lower grades? Can these dangers be safeguarded? What types of reading matter do you now read orally outside the schoolroom? Are these the types which your pupils are asked to read orally?
5. What are the circumstances under which you last read aloud? Do your pupils have the same incentives for reading clearly and interestingly that you had on that occasion?
6. What are some of the things you do to assist your pupils in developing ability to comprehend the meaning of the printed page? Do you know of faulty habits which some of them have which prevent their centering attention upon the meaning? Do you know which pupils read with accuracy? Which ones read rapidly?
7. Which of the silent reading tests described in this chapter would you select for use in the intermediate grades? Why? Which ones for use in the seventh and eighth grades and the high school? Why?
8. Is a power test more nearly in agreement with our educational objectives in reading than in arithmetic? Why?
9. How long does it take you to become familiar with the reading difficulties of each child when you receive a new class of, say, thirty children? Would you consider it economical if some tests were available by means of which you could discover these difficulties as well as others the first day and thus prepare a chart of each child's instructional needs? How long at the beginning of a term could you afford to spend in making such a diagnosis?

SELECTED BIBLIOGRAPHY

- Anderson, C. J., and Merton, Elda. "Remedial Work in Reading"; in *Elementary School Journal*, vol. 20, pp. 685-701, 772-91. (May and June, 1920.)
- Anderson, Homer Willard. *Measuring Primary Reading in the Dubuque Schools*. The Harris-Anderson Tests. (Dubuque, Iowa, 1916.)
- Bonser, Frederick G., and others. "Vocabulary Tests as Measures of School Efficiency"; in *School and Society*, vol. 2, pp. 713-18. (November 13, 1915.)

- Burgess, May Ayres. *The Measurement of Silent Reading*. (New York: Russell Sage Foundation, 1921. 163 pp.)
- Burgess, May Ayres. "Classroom Grouping for Silent-Reading Drill"; in *Elementary School Journal*, vol. 22, pp. 269-78. (December, 1921.)
- Courtis, S. A. "Standard Rates of Reading"; in *Fourteenth Yearbook of the National Society for the Study of Education*, part 1, pp. 44-59. (Bloomington, Illinois, Public School Publishing Company, 1915.)
- Courtis, S. A. *Measurement of Classroom Products: Survey of the Gary Public Schools*. (New York: General Education Board, pp. 279-91, 324-31.)
- Daley, H. C. "Equivalence of Forms I and II of the Burgess Picture Supplement Scale for Measuring Silent Reading Ability"; in *Journal of Educational Research*, vol. 4, pp. 71-72. (June, 1921.)
- Gates, Arthur I. "An Experimental and Statistical Study of Reading and Reading Tests"; in *Journal of Educational Psychology*, vol. 12, pp. 303-14, 378-91, 445-64. (September, October, November, 1921.)
- Gates, Arthur I. "A Study of Reading and Spelling with Special Reference to Disability"; in *Journal of Educational Research*, vol. 6, pp. 12-24. (June, 1922.)
- Gray, Clarence Truman. *Types of Reading Ability as Exhibited through Tests and Laboratory Experiments*. Supplementary Educational Monograph no. 5. (Chicago: University of Chicago, 1917.)
- Gray, Clarence Truman. *Deficiencies in Reading Ability: Their Diagnosis and Remedies*. (New York: D. C. Heath and Company, 1922. 420 pp.)
- Gray, William Scott. "A Coöperative Study of Reading in Eleven Cities in Northern Illinois"; in *Elementary School Journal*, vol. 17, pp. 250-65. (December, 1916.)
- Gray, William Scott. "A Study of the Emphasis on Various Phases of Reading Instruction in Two Cities"; in *Elementary School Journal*, vol. 17, pp. 178-86. (November, 1916.)
- Gray, William Scott. "Methods of Testing Reading"; in *Elementary School Journal*, vol. 16, pp. 231-46, 281-98. (January and February, 1916.)
- Gray, William Scott. *Studies of Elementary School Reading through Standardized Tests*. Supplementary Educational Monographs, vol. 1, no. 1. (Chicago: University of Chicago, 1917. 157 pp.)
- Gray, William Scott. *A Coöperative Study of Reading in Sixteen Cities in Indiana*. Indiana University Studies no. 39. (Bloomington, Indiana: University of Indiana, 1918.)
- Gray, William Scott. "The Use of Tests in Improving Instruction"; in *Elementary School Journal*, vol. 19, p. 121. (October, 1918.)
- Gray, William Scott. "Reading in the Elementary Schools of Indianapolis"; in *Elementary School Journal*, vol. 19, pp. 336-53, 419-44, 506-31. (January, February, March, 1919.)
- Gray, William Scott. "Value of Informal Tests of Reading Accomplishment"; in *Journal of Educational Research*, vol. 1, pp. 103-11. (February, 1920.)

152 EDUCATIONAL TESTS AND MEASUREMENTS

- Gray, William Scott. "Diagnostic and Remedial Steps in Reading"; in *Journal of Educational Research*, vol. 4, pp. 1-15. (June, 1921.)
- Gray, William Scott. "Individual Difficulties in Silent Reading in the Fourth, Fifth, and Sixth Grades"; in *Twentieth Yearbook of the National Society for the Study of Education*, part II, pp. 39-53. (Bloomington, Illinois: Public School Publishing Company, 1921.)
- Gray, William Scott. *Remedial Cases in Reading: Their Diagnosis and Treatment*. Supplementary Educational Monograph no. 22. (Chicago: University of Chicago, 1923. 208 pp.)
- Greene, Harry A. "Measuring Comprehension of Content Material"; in *Twentieth Yearbook of the National Society for the Study of Education*, part II, pp. 114-26. (Bloomington, Illinois: Public School Publishing Company, 1921.)
- Haggerty, M. E. "Scales for Reading Vocabulary of Primary Children"; in *Elementary School Journal*, vol. 17, pp. 106-15. (October, 1916.)
- Haggerty, M. E. *The Ability to Read: Its Measurement and Some Factors Conditioning it*. University of Indiana Studies no. 34. (Bloomington: University of Indiana, 1917.)
- Henmon, V. A. C. "Improvement in School Subjects throughout the School Year"; in *Journal of Educational Research*, vol. 1, pp. 81-95. (February, 1920.)
- Jones, E. E., and Lockhart, A. V. "A Study of Oral and Silent Reading in the Elementary Schools of Evanston"; in *School and Society*, vol. 10, pp. 587-90. (November 15, 1919.)
- Judd, C. H. "Reading"; in *Fifteenth Yearbook of the National Society for the Study of Education*, part I, pp. 111-19. (Bloomington, Illinois: Public School Publishing Company, 1916.)
- Judd, Charles H. *Measuring the Work of the Public Schools*. Report of the Survey Committee of the Cleveland Foundation. (Cleveland, Ohio: The Survey Committee of the Cleveland Foundation, 1916. 290 pp.)
- Kallom, A. W. "Reproduction as a Measure of Reading Ability"; in *Journal of Educational Research*, vol. 1, pp. 359-68. (May, 1920.)
- Kelley, T. L. "Thorndike's Reading Scale Alpha 2, adapted to Individual Testing"; in *Teachers College Record*, vol. 18, pp. 253-60. (May, 1917.)
- Kelly, F. J. "The Kansas Silent Reading Test"; in *Journal of Educational Psychology*, vol. 7, pp. 63-80. (February, 1916.)
- Lloyd, S. M., and Gray, C. T. *Reading in a Texas City, Diagnosis and Remedy*. University of Texas Bulletin no. 1853, Education Series no. 4, pp. 48-62. (Austin: University of Texas, 1918.)
- McCall, William A. "Proposed Uniform Method of Scale Construction"; in *Teachers College Record*, vol. 22, pp. 31-51. (January, 1921.)
- McLeod, L. S. "The Influence of Increasing Difficulty of Reading Material upon Rate, Errors, and Comprehension in Oral Reading"; in *Elementary School Journal*, vol. 18, pp. 523-32. (March, 1918.)
- Mead, Cyrus D. "Silent Reading *versus* Oral Reading with One Hundred

- Sixth-Grade Pupils"; in *Journal of Educational Psychology*, vol. 7, pp. 63-80. (February, 1916.)
- Monroe, Walter S. "A Report on the Use of the Kansas Silent Reading Tests with over One Hundred Thousand Children"; in *Journal of Educational Psychology*, vol. 9, pp. 600-08. (December, 1917.)
- Monroe, Walter S. "Monroe's Standardized Silent Reading Tests"; in *Journal of Educational Psychology*, vol. 9, pp. 303-12. (June, 1918.)
- Monroe, Walter S. "The Law of the Single Variable"; in *Journal of Educational Research*, vol. 4, pp. 58-60. (June, 1921.)
- Monroe, Walter S. *The Illinois Examination*. University of Illinois Bulletin, vol. 19, no. 9, Bureau of Educational Research Bulletin no. 6. (Urbana: University of Illinois, 1921. 70 pp.)
- Monroe, Walter S. *A Critical Study of Certain Silent Reading Tests*. University of Illinois Bulletin, vol. 19, no. 22, Bureau of Educational Research Bulletin no. 8. (Urbana: University of Illinois, 1922. 52 pp.)
- Otis, Arthur S. "Considerations Concerning the Making of a Scale for the Measurement of Reading Ability"; in *Pedagogical Seminary*, vol. 23, pp. 528-49. (December, 1916.)
- Peters, C. C. "The Influence of Speed Drills upon the Rate and Effectiveness of Silent Reading"; in *Journal of Educational Psychology*, vol. 8, pp. 350-66. (June, 1917.)
- Pressey, S. L., and Pressey, L. W. "The Relative Value of Rate and Comprehension Scores in Monroe's Standardized Silent Reading Tests as Measures of Reading Ability"; in *School and Society*, vol. 11, pp. 747-49. (June 19, 1920.)
- Pressey, Luella W. *Reading Scales for the Second, Third, and Fourth Grades*. University of Indiana Extension Division Bulletin, vol. 6, no. 12, pp. 46-52. (Bloomington: University of Indiana, 1921.)
- Pressey, L. W., and Pressey, S. L. "A Critical Study of the Concept of Silent Reading Ability"; in *Journal of Educational Psychology*, vol. 12, pp. 25-31. (January, 1921.)
- Pressey, Luella C. "A First Report on Two Diagnostic Tests in Silent Reading for Grades II to IV"; in *Elementary School Journal*, vol. 22, pp. 204-11. (November, 1921.)
- Richards, Alva M., and Davidson, Percy E. "Correlations of Single Measures in Some Representative Reading Tests"; in *School and Society*, vol. 4, pp. 375-77. (September 2, 1916.)
- Starch, Daniel. "The Reliability of Reading Tests"; in *School and Society*, vol. 8, pp. 86-90. (July 20, 1918.)
- Theisen, W. W. "Factors Affecting Results in Primary Reading"; in *Twentieth Yearbook of the National Society for the Study of Education*, part II, pp. 1-24. (Bloomington, Illinois: Public School Publishing Company, 1921.)
- Theisen, W. W. "Does Intelligence Tell in First-Grade Reading?" in *Elementary School Journal*, vol. 22, pp. 530-34. (March, 1922.)

154 EDUCATIONAL TESTS AND MEASUREMENTS

- Thorndike, Edward L. "The Measurement of Ability to Read"; in *Teachers College Record*, vol. 15, no. 4. (September, 1914.)
- Thorndike, Edward L. "An Improved Scale for Measuring Ability in Reading"; in *Teachers College Record*, vol. 16, pp. 445-67; vol. 17, pp. 40-67. (November, 1915; January, 1916.)
- Thorndike, Edward L. "The Measurement of Achievement in Reading: Word Knowledge"; in *Teachers College Record*, vol. 17, pp. 430-54. (November, 1916.)
- Thorndike, Edward L. "Reading as Reasoning: A Study of Mistakes in Paragraph Reading"; in *Journal of Educational Psychology*, vol. 8, pp. 323-32. (June, 1917.)
- Thorndike, Edward L. "The Understanding of Sentences: A Study of Errors in Reading"; in *Elementary School Journal*, vol. 18, pp. 98-114. (October, 1917.)
- Uhl, W. L. "The Use of the Results of Reading Tests as Bases for Planning Remedial Work"; in *Elementary School Journal*, vol. 17, pp. 266-75. (December, 1916.)
- Updegraff, Harlan, and King, LeRoy A. "Second Annual Report of the Bureau of Educational Measurements"; in *Sixth Annual Schoolmen's Week Proceedings*. (Philadelphia: University of Pennsylvania, 1919.)
- Updegraff, Harlan, and King, LeRoy A. "Third Annual Report of the Bureau of Educational Measurements"; in *Seventh Annual Schoolmen's Week Proceedings*. (Philadelphia: University of Pennsylvania, 1920.)
- Wassen, H. W. "Report of an Experiment in the Use of the Kansas Silent Reading Test with Korean Students"; in *Educational Administration and Supervision*, vol. 3, pp. 98-101. (February, 1917.)
- West, Paul V. "The Monroe Silent Reading Test," in *School and Society*, vol. 13, p. 510. (April 23, 1921.)
- Wilson, Estaline. "Specific Teaching of Silent Reading"; in *Elementary School Journal*, vol. 22, pp. 140-46. (October, 1921.)
- Witham, E. C. "Scoring the Monroe Standardized Silent Reading Tests"; in *Journal of Educational Psychology*, vol. 9, p. 516. (November, 1918.)
- Wyman, J. Benson, and Wendle, Miriam. "What is Reading Ability?" *Journal of Educational Psychology*, vol. 12, pp. 518-31. (December, 1921.)
- Ziedler, Richard. "Tests in Silent Reading in the Rural Schools of Santa Clara County, California"; in *Elementary School Journal*, vol. 18, pp. 55-62. (September, 1916.)

CHAPTER IV

HANDWRITING

I. THE PROBLEM OF THE MEASUREMENT OF HANDWRITING

The traits measured in handwriting. There are several ways of measuring handwriting. When copy-books represented our objectives the script produced by children was constantly compared with, that is, measured by, these copy-book specimens. Consequently the emphasis was placed on well-shaped letters and pretty lines, and rate of writing and legibility were largely disregarded as objectives. Teachers who are watching the writer frequently attend to the writer's position, movement, and apparent ease of production. In ordinary situations outside the schoolroom, we commonly judge handwriting according to its degree of legibility. If we read it easily, we give it a high rating. If we have difficulty in reading it, we call it poor. In case the handwriting is conspicuous for its beauty or the lack of it, we may take beauty into account in our judgment. In the refinement of our methods of measuring handwriting, we have focused our attention upon legibility and beauty. The combination of these characteristics we call quality. Measurement as it relates to handwriting is concerned with both the act of writing and the script which is produced. For diagnostic measurement each of these factors should be analyzed into more simple and more readily measurable adjustments and quantities.

Difficulties encountered in measuring handwriting. Here, as in the analysis of other school subjects, we may find characteristics which are easily measured, but of so little value as to make measurement unprofitable. We shall also

find some traits which are measured with difficulty and without precision, but which must be measured because of their worth. Many teachers seem to be concerned with the movement which the pupils are using. This could be accurately measured, but it is not of sufficient importance, in the ordinary classroom situation, to warrant the time and effort required to so measure it. Undue emphasis on movement has doubtless robbed legibility of the attention it deserves. On the other hand, legibility cannot be measured as precisely and objectively as movement, but even a rough measure of legibility aids in calling attention to it as a goal.

The mention of legibility as a characteristic of handwriting suggests the difficulty encountered in its measurement. It is easy to determine whether the spelling of a word or the answer to an example in arithmetic is right or wrong. It is not so easy to judge whether a specimen of handwriting is legible or illegible. Instead of the clear-cut distinction between correct and incorrect, we here find need for recognition of degrees of legibility. These cannot be easily defined. A specimen of handwriting which is read with great difficulty by one person may be read rapidly by another. Legibility depends in part on the skill of the reader. Legibility is not a simple characteristic. Size of letters, kind of line, spacing, and other factors affect it. Legibility concerns the reader. Rate of writing and ease of performance in writing concern the writer.

II. MEASURING THE ACT OF HANDWRITING

Discussions of the measurement of handwriting have usually emphasized measurement of the script produced. However, Judd¹ has asserted, "The habit of writing is not to be found in lines on paper — it is in human beings." Although

¹ Judd, Charles Hubbard, *Genetic Psychology for Teachers*, chap. vi. (Chicago: D. Appleton and Company, 1903.)

measurement of the script furnishes an indirect index of the act of writing and may be made with a minimum of effort and of complexity of technique, it needs to be supplemented and checked by more direct measurements of the act of writing. Such measurements are especially valuable for the laboratory and clinical worker.

Extreme emphasis upon position in handwriting not justified. Much has been said concerning the hygiene of position and movement in handwriting. Freeman¹ gives the warning that incorrect position assumed in writing results in spinal curvature, eye-strain, crowding of the lungs and digestive organs, and other disasters. An examination of the authorities quoted by this writer produces a doubt as to the validity of his conclusion for our school children. Cause and effect may have been confused. Consultation with three reputable practicing orthopedists brought the unanimous opinion that the position taken during the handwriting period could have very little effect on bone growth and symmetry. American children spend very little time in writing, and that usually relatively late in childhood after bones

¹ Freeman, Frank N., *The Teaching of Handwriting*, pp. 32-55. (Boston: Houghton Mifflin Company, 1914.)

Freeman, Frank N., "Principles of Method in Teaching Writing as Derived from Scientific Investigation"; in *Eighteenth Yearbook of the National Society for the Study of Education*, part II, pp. 11-25. (Bloomington, Illinois: Public School Publishing Company, 1919.)

"Bending and turning the trunk, which causes curvature of the spine, may be largely avoided by requiring the pupil to face the desk squarely, etc.

"The remaining defect of posture consists in turning or bending the head. The danger is, of course, again that curvature of the spine will result from a constant holding of the head in any but an erect position.

"Rule I. The writer should face the desk squarely.

"Evidence. Statistical evidence indicates that a side position causes spinal curvature and eye-strain.

"Rule II. Both forearms should rest on the desk for approximately three quarters of their length.

"Evidence. Statistical evidence indicates that when one elbow is unsupported spinal curvature is produced."

have considerable rigidity. Postures assumed in early childhood and positions in which the infant is placed are of more importance. Furthermore, observation of schoolroom conditions supports the assertion that postures taken in penmanship practice have only small influence on postures taken in writing at other times. Discussions of bodily posture, hygiene, and physiology of handwriting serve to show the loose and unscientific treatment of this phase of measurement of handwriting. There is need of encouragement of correct posture on the part of children in all school situations. There is equally great need of insistence on correct postures in home and playground situations. Overemphasis on posture in the limited period of penmanship practice and supervised writing under the fear of curvature of the spine and the like, would seem to be more irritating than constructive. The statements concerning posture during handwriting seem to be a survival from the period when children sat in "position." In the modern school there is such freedom of movement that this discussion is no longer significant.

There are no devices for the direct measurement of posture. It is conceivable that a score card could be devised assigning points on the basis of an ideal posture.

Measurement of handwriting movement. Judd's¹ hand-tracer and its modification devised by Freeman² and Nutt³ give accurate and valid measures of the kind of movement and its amplitude. These devices have in common a stylus or pencil fastened to the wrist or hand which traces on a paper the path of movement of the hand during the writing. If

¹ Judd, Charles Hubbard, *Genetic Psychology for Teachers*. (Chicago: D. Appleton and Company, 1903.)

² Freeman, F. N., *Yale Psychological Studies*, New Series, vol. II, no. 1; *Psychological Monographs*, XVII, 1914.

³ Nutt, H. W., "Rhythm in Handwriting"; in *Elementary School Journal*, vol. 17, pp. 432-45. (February, 1917.)

there is no finger-movement, this stylus traces almost an exact copy of the specimen written by the pen which is held between the thumb and index finger. If there is complete finger-movement and very little arm or "muscular" movement, this stylus traces nearly a straight line. The use of these devices shows rather conclusively that the estimates of the amounts of finger- or arm-movement usually made by teachers are very much in error. The valid and reliable devices for the measurement of form of performance are limited to the hand-tracers as devised by Judd and modified by Freeman and Nutt.

Results of measurement of movement. The use of the hand-tracers seems to have been limited to a few investigators. In view of the exceedingly optimistic claim of the advantages of muscular movement made by the champions of the several commercialized systems of penmanship, it would seem that some of them could have put their claims to this acid test of scientific investigation. Typical results secured by use of the hand-tracer are given in the following quotation from Judd's *Genetic Psychology for Teachers*.¹

In ordinary writing the fine formative movements are executed by the fingers; the movements which carry the fingers forward are executed by the hand or arm; and the pauses between groups of letters are utilized for longer forward arm movements which bring the hand back into an easy working position. . . . Each individual has his own peculiar coördination of arm and hand and finger movements. . . . The forms of coördination are as numerous and as various as are the individuals who write.

Any change in the conditions under which the subject writes will modify the character of coördination. A change from a hard pencil to a soft pencil, or a change from a vertical position of the paper to an oblique position will be sufficient to produce noticeable variations in the character of the muscular coördination, even when the product of the movement — that is, the written letters — conform very closely to the same type.

¹ Chapter VI.

Nutt¹ using a complicated form of the hand-tracer, secured results which have an important bearing upon considerations of the kind of movement that should be taught. He found that arm-movement is closely correlated with age. "None of the systems of penmanship develop any appreciable degree of arm-movement in the younger children." The systems which were used in the schools studied were the Palmer system, the Ransomalian system, the Kansas State copy-book, and the Berry copy-book. From ten years on some children adopt an arm-movement and some do not. With the facts that arm-movement is not correlated with quality, rhythm, or speed, the question arises, What is the value of arm-movement to handwriting? Since it develops with age and since children mature at varying rates, it would seem that Nutt's conclusion that arm-movement should be left to individual selection and development is warranted. There are some teachers who believe that arm-movement must be developed. They should postpone active training until the pupils are ten or more years of age except in

¹ "The purpose of the apparatus was to secure a specimen of the writer's handwriting, a record of the amount of arm-movement employed in writing the copy, and a record of the movement of the writer's pencil so spread out from left to right upon a strip of paper moving at a known speed as to make possible the measuring of the time taken for the writing of each stroke. The mechanism of the apparatus afforded a means of measuring speed by providing underneath the sheet of paper on which the subject wrote a typewriter ribbon belt moving from right to left and directly beneath the ribbon a moving strip of paper, on which a traced record was received. Thus, when the writer wrote upon the stationary sheet of paper the path of the pencil point was made upon the moving strip of paper. The speed of the strip of paper was recorded by tracing on it the path of an electrically driven vibrating pen. This path is called the time line.

"The arm-movement was recorded by attaching a system of levers to the back of the hand . . . through which the movement of the hand was transmitted to a pen writing on another sheet of paper. The mechanism was kept so accurately adjusted that an almost perfect reproduction of the written copy was secured if the writer was using no finger-movement." (*Op. cit.*, pp. 433-34.)

the few cases in which the arm-movement is spontaneously adopted by younger children.

Rhythm, speed, and age are closely correlated. This means that standards of speed must be adjusted to age levels. Since rhythm and quality are not closely related, it is quite probable that the persistence of certain rhythm is detrimental to quality. Nutt¹ observes that rapid writers tend to slur over the parts of letters which interrupt their characteristic rhythms. The reconciliation of letter forms to the rhythmic movements is a critical point which requires further investigation. Once the best forms of letters are determined, the proper drills must be prepared.

There is need for a thorough study of the rhythmic movements of children's writing throughout all the age levels of childhood, from the period when children draw to the adult stage of rapid rhythmic writing. Paralleling this study there should be a study of the letter forms which are best suited to the typical rhythms. When we face these problems we wonder if communication will not take on new forms before these investigations make a scientifically directed training in handwriting possible. With typewriting and stenography taught everywhere, with cheap portable typewriters, stenotype machines, the dictaphone, and possibilities yet undeveloped in the radio, it appears possible that highly trained handwriters may not be desirable. History indicates that man is more fond of altering his environment than he is of developing his own responses, when this development does not have immediate satisfactions greater than those attendant on well-developed penmanship.

III. MEASUREMENT OF THE RATE OF HANDWRITING

General procedure. The rate of handwriting can be measured most conveniently by having all of the children of a

¹ *Op. cit.*

class write a memorized copy for a specified time. If they all write the same copy the number of letters written can easily be counted. To obtain an accurate and reliable measurement of rate the conditions under which the writing is done should be carefully controlled. (See page 179.) The teacher should guard against interruptions and delays by making sure that all pupils are provided with pen, ink, and paper. Since one of the purposes of measuring rate is that of making comparisons with the record of other similar groups and with standard scores, the teacher should make certain that conditions are the same in her classroom as those under which samples of handwriting were gathered which furnished the scores for comparison. The most generally accepted conditions are as follows:

Time. The time should be two or three minutes. With younger pupils three minutes is not too long. Older children will write for two minutes at approximately the same rate that they will for a longer time. The labor of counting the letters written is increased when they write for a longer time. Exactly the time agreed upon should be allowed. The teacher should use a stop-watch, or, if using an ordinary watch, she should record the exact position of the second hand when the signal to start is given in order to avoid error.

Copy. Pupils should be asked to write a suitable selection which they have memorized. To guard against lapses of memory the pupils should be asked to repeat in concert the selection which is to be used. If convenient provide each pupil with a type written or printed copy of it. When this cannot be done the selection may be written on the blackboard where all can see it. The selection should contain no words which the pupils cannot spell readily. If there are words which seem to present spelling difficulties, have the children write them a few times before the test is begun. Material which the pupils must compose as they

write is useless for measuring the rate of writing. The rate of writing unfamiliar material from a printed copy will vary with the pupil's rate of reading and will not give a true measure of his rate. Dictated material should be used only when the teacher wishes to control the rate, not when it is to be measured. The selections which have been used vary considerably. If the teacher is planning to use one of the handwriting scales for measuring quality, it is well to use the selection which is presented in the scale. Thus, for the Thorndike Scale use, "Then the carelessly dressed gentleman stepped lightly into Warren's carriage and held out a small card. John vanished behind the bushes and the carriage moved along down the driveway." If the Ayres Handwriting Scale, Gettysburg Edition, is to be used, have the pupils write the first three sentences of Lincoln's *Gettysburg Address*. Freeman's scales make use of a sentence in which every letter in the alphabet occurs one or more times: "A quick brown fox jumps over the lazy dog."

Directions to pupils. The instructions which are given to the pupils are a very strong influence in determining the rate and the quality of the writing. They are so important that great care should be exercised in using an exact formula. There is considerable evidence on this point. Sackett¹ showed that students who knew they were being tested wrote slightly slower and a better quality than they ordinarily wrote. Freeman showed that the call for quality reduced the rate 3.7 per cent and improved the quality 6.2 per cent while the call for rapid writing decreased the quality 9.1 per cent and increased the speed 27.2 per cent. The writer found in a class of thirty-seven college students that the rate showed an average increase of 43.9 per cent when the instructions emphasized rate, but writing for the best quality re-

¹ Sackett, L. W., "Comparable Measures of Handwriting"; in *School and Society*, vol. 4, pp. 640-46. (October 21, 1916.)

duced the rate 15.3 per cent from the rate of ordinary writing. Hence, a complete and satisfactory measurement of rate should include three specimens, one (the first), written in response to instructions calling for the ordinary rate of writing, another written at maximum rate at which writing can be made legible, and a third written for the best quality regardless of rate. Freeman thinks the best single set of samples could be secured in response to the instructions — “Write as well as you can and as rapidly as you can.”

IV. MEASUREMENT OF QUALITY OF HANDWRITING

Thorndike's Handwriting Scale. As stated above, “quality” is a general term including legibility, beauty, and several other factors. This general characteristic is measured by means of scales, score cards, and certain combinations of these. Thorndike¹ constructed the first scale for the measurement of the quality of handwriting. It consists of a series of specimens of handwriting arranged in order of increasing quality. These specimens were assigned values on the basis of three characteristics: beauty, legibility, and general merit. The degree of these characteristics represented in the specimens of the scale was determined by the consensus of opinion of competent judges.² The values assigned to the specimens of the scale range from 4 to 18, one or more specimens being displayed for each step of the scale. Three steps of the scale are shown in Fig. 7. The scale is printed on a large sheet of paper.

Teachers and expert penmen have made three criticisms of this scale: (1) The form of the test makes it inconvenient to use; it is too large, and the square shape is not so conven-

¹ Thorndike, E. L., “Handwriting”; in *Teachers College Record*, vol. 2, no. 2. (March, 1910.)

² For a description of the procedure of constructing a scale of this type see Monroe, Walter S., *An Introduction to the Theory of Educational Measurements*, chap. vi. (Boston: Houghton Mifflin Company, 1923.)

Quality 11.

John vanished behind the
bushes and the carriage
moved along down the
driveaway. The audience

driveaway. The audience of passers-by, which
had been gathering about them melted away
in an instant leaving only a poor old lady on
the curb. Albert was sadly striking

Quality 9.

Then the carelessly dressed gentlemen
stepped lightly into Warren's carriage and
held out a small card, John vanished behind the
by which had been gathering about them melt-
ed away in an instant leaving only a poor
old lady on the curb. Albert was sadly

Then the carelessly dressed gentleman
stepped lightly into Warren's carriage moved
and held out a small card, John vanished

Quality 8.

Moved along down the driveaway. The
audience of passers-by which had
been
gathering about them melted away.

Then the carelessly gentleman step-
ped lightly into Warren's carriage and
held out a small card, John vanished be-
hind the bushes and the carriage moved

FIG. 7. A SECTION OF THE THORNDIKE HANDWRITING SCALE
(Reduced one half in size.) Quality 9 of this scale is approximately equal to quality
40 of the Ayres Scale. Quality 11 is better than the Ayres 50.

ient as the rectangular form used by Ayres. (2) The specimens in the scale do not represent the types of penmanship now generally taught. This objection is raised, especially by the exponents of certain systems of penmanship. (3) The steps of the scale are represented by varying numbers of specimens; some steps have four specimens and some only one. In further criticism Starch¹ has shown that the specimens which are assigned the same value in the Thorndike Scale do not have the same value when rated by a larger number of judges than Thorndike used, or when rated on Starch's new scale.

Although all of these criticisms have weight, it must be remembered that the Thorndike Handwriting Scale was the first quality scale constructed and that most of the succeeding workers have profited by Thorndike's pioneer work. Inasmuch as the reliability of the measurement of the quality of handwriting depends largely on the skill of the person using the scale, it is probable that a well-trained judge can use the Thorndike Scale with very good results.

The Ayres Handwriting Scales. Ayres² has constructed three scales for the measurement of the quality of handwriting, the Three Slant Scale, the Adult Scale, and the Gettysburg Edition. He attempted to evaluate the specimens of handwriting selected for his scales upon the single basis of legibility. The method of construction differs from that followed by Thorndike.

The Three Slant Scale consists of three types of specimens of handwriting, vertical, semi-slant, and full slant. Each of

¹ Starch, D., "A Scale for Measuring Handwriting"; in *School and Society*, vol. 9, pp. 154-58 and 184-88. (February 1 and February 8, 1919.)

² Ayres, L. P., *A Scale for Measuring the Handwriting of School Children*. Russell Sage Foundation Bulletin no. 113. (New York: Russell Sage Foundation.)

Ayres, L. P., *A Scale for Measuring the Handwriting of Adults*. Russell Sage Foundation Bulletin no. 138. (New York: Russell Sage Foundation.)

50	60
Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived	Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal Now we are engaged in a great civil war, test- ing whether that nation, or any nation so con- ceived and so dedicated

FIG. 8. TWO SECTIONS FROM THE AYRES HANDWRITING SCALE, GETTYSBURG EDITION

these three types is represented by eight degrees of quality to which are assigned the numerical values, 20, 30, 40, up to 90. In using this scale it must be remembered that these values are not the same as the per cents used in reporting "grades."

The Adult Scale is similar to the Three Slant Scale in its general plan. The specimens are taken from samples of adults' handwriting, which were rated by trained judges using the Three Slant Scale.

The Gettysburg Edition (Fig. 8) has the same steps of value as the Three Slant Scale, but has only one specimen of medium slant writing for each step of the scale. Of this scale Ayres says, "The purpose of the changes introduced in the present edition is to increase the reliability of measurement of handwriting through standardizing methods of securing and scoring samples and through making numerous improvements in the scale itself designed to reduce variability in results secured through its use." Ayres added to this scale instructions for securing and scoring samples of handwriting and graphical representations of standards.

The Ayres Gettysburg Edition has escaped most of the practical criticisms which were made concerning the Thorndike Handwriting Scale. Some penmen say it should have specimens of poorer quality than sample 20, and specimens better than sample 90. This criticism is based on needs which seldom arise. The value of 20 implies that 0 quality would be represented by a specimen two steps lower on the scale. The value of 90 suggests that there is room for better specimens. In case a specimen of handwriting appears to be poorer than 20, the judge can assign a value of 15, 10, 5, or 0. Likewise a sample which is judged better than 90 can be assigned values of 95, 100, 105, or even more. Starch¹ found that the values assigned to the specimens were open to the

¹ Starch, D., "A Scale for Measuring Handwriting"; in *School and Society*, vol. 9, pp. 154-58, 184-88. (February 1 and February 8, 1919.)

same criticism as those of the Thorndike Scale. He says, "It is obvious that the units of the scale, when expressed in terms of general merit, are so uneven that the validity of the scale is seriously impaired." Starch and others have pointed out that Ayres evaluated the specimens of his Three Slant Scale on the basis of the average rate at which they were read, but that the scale is used to measure general merit rather than legibility. As in the case of the Thorndike Scale these criticisms need not be interpreted as a condemnation of the scales. The final test of the value of these handwriting scales is the consistency of the ratings which are made by judges using them. This question will be discussed under the head of reliability.

Because of its convenient form, its availability and the instructions which accompany it, the Ayres Handwriting Scale, Gettsyburg Edition, has become the most popular and most widely used handwriting scale. It is now doubtless the most usable of all scales for the beginner in educational measurements.

Starch Handwriting Scale. Starch¹ has prepared a handwriting scale which seems to be free from certain of the defects of the other scales. It is printed on a cardboard, fourteen by forty-five inches. This inconvenient size is the most serious objection to the scale. Great care was used in selecting the specimens of handwriting which are representative of a desirable type of penmanship. The original collection of 627 samples was reduced by the judgment of an expert penman to 227 samples which included as the best samples the writing of expert penmen. Certain refinements were made in Thorndike's method of construction which tend to make the evaluation of the specimens more accurate. This increases the accuracy of the measurements. Other advan-

¹ Starch, D., "A Scale for Measuring Handwriting"; in *School and Society*, vol. 9, pp. 154-58, 184-88. (February 1 and February 8, 1919.)

tages are claimed for this scale. It includes uniform text material, including capital letters. There is a continuous series of steps from "quality 0" to "quality 20." The better specimens of the scale are in a business style of handwriting. The scale is printed in blue ink to make the samples resemble pupils' ordinary writing. Norms of attainment in both rate and quality are given on the scale. A simple plan is also given for converting the measures into ordinary school marks.

The New York City Penmanship Scale. This scale was prepared by Lister¹ and Myers of the Brooklyn Training School for Teachers. They followed the method originated by Thorndike. It is in reality three scales: one for form of letter formation, one for spacing, and one for movement as judged from the quality of the line. The scale is printed on pages 24 by 26 inches and for this reason is inconvenient to use. In describing this scale, Nifenecker² gives tentative standards for each characteristic measured. He also gives the average rate at which each degree of quality was written by pupils in the grades from the fourth through the eighth. This scale is made up of samples of the handwriting of the children of the New York City schools.

Freeman Handwriting Scale.³ This scale is in reality five scales, one for each of the following characteristics of handwriting: uniformity of slant, uniformity of alignment, equality of line, letter formation, and spacing. These five scales are now printed on one sheet of paper or chart, and each

¹ Lister, Clyde C., *The New York City Penmanship Scale*. Brooklyn Training School for Teachers Bulletin no. 3. (Brooklyn, New York: Brooklyn Training School for Teachers, 1919.)

² Nifenecker, E. A., "Grade Norms for the New York City Penmanship Scale"; in *Journal of Educational Research*, vol. 2, p. 809. (December, 1920.)

³ Freeman, F. N., *The Teaching of Handwriting*. (Boston: Houghton Mifflin Company, 1915.) Also, "An Analytical Scale for the Judging of Handwriting"; in *Elementary School Journal*, vol. 15, p. 432. (April, 1915.)

scale is called a division. The first of the five divisions of the Freeman Scale represents three degrees of uniformity of slant. In using this division, as in using the next division, judgments will be made more easily if a slant and alignment gauge is used.¹ The second division represents uniformity of alignment. The user must be careful to note that letters which are close together show deviations in alignment more prominently than letters written farther apart.

The third division shows the quality of line or stroke. A reading-glass will aid in judging with this division. The fourth division is intended to measure letter formation. Freeman describes eight illegible forms of letters which should be counted as errors. Two principles should control here: first, whatever slant or type of script the pupil may use, consistency to that choice should be maintained; and, second, no letter should vary from its recognized form so much as to be easily mistaken for another letter. The fifth division shows different kinds of spacing. Letters may be crowded together or spread too far apart. The same applies to words.

In each division the three degrees of excellence are given scores of 1, 3, and 5 respectively. The intermediate values of 2 and 4 may also be used. If the old edition of the scale is used, the scores assigned to the specimens of letter formation are 2, 6, and 10. Freeman suggests that the specimens be scored by using the score for letter formation as placed on the new edition of the chart, and then doubling these scores in making up the total score.

¹ Freeman, F. N., *The Teaching of Handwriting*, p. 151. The slant gauge consists of three rows of parallel lines. The lines in one row are vertical and in each of the other rows the lines are set at a uniform slant. The alignment gauge consists of one straight line four or five inches long. These lines may be drawn on transparent paper and placed over a specimen of handwriting to assist in determining the deviations from uniformity in slant and alignment.

The score card for detailed analysis of handwriting. The score card represents another attack upon the problem of measurement in this field. Such instruments as the Ayres and the Thorndike scales do not require that the user make an analysis of general merit or quality. The score card requires that each of the essential elements of handwriting be judged separately and assigned a value. The score card devised by Gray¹ weights the value of each of the essential elements of handwriting so that the highest value which can be assigned to slant is 5, while spacing of letters may receive 18, neatness, 13, etc. (See Fig. 9.) The total of the values assigned to the separate elements may be used as a measure of general merit. So far there is no evidence to show that a measure of general merit obtained in this way will be more accurate than the measure obtained by the use of any one of the general scales. Some claim that in the Gray Score Card the elements of handwriting have not been correctly weighted. However, it has the advantage that its use trains the user in the analysis of handwriting and directs attention to the essential elements. Gray well defends the device by saying that agriculturists have long used such score cards to secure very satisfactory and accurate results in judging grain and live stock.

Objective indices of legibility. Several forms of an "index of legibility" have been proposed by certain investigators. Both Witham² and Sackett³ have proposed that the specimens which are being rated be read and the time of

¹ Gray, C. Truman, *A Score Card for the Measurement of Handwriting*, University of Texas Bulletin no. 37. (Austin: University of Texas, 1915.)

² Witham, Ernest, "The Most Accurate Measure of Handwriting"; in *Journal of Educational Administration and Supervision*, vol. 6, pp. 150-58. (June, 1920.) Also earlier article in *School Board Journal*, May, 1914, and in *Journal of Educational Administration and Supervision*, May, 1915.

³ Sackett, L. W., "Comparable Measures of Handwriting"; in *School and Society*, vol. 4, pp. 640-45. (October 21, 1916.)

Pupil..... Age..... Date.....
 Grade..... School.....
 Sample Number..... Teacher.....

SAMPLE	PERFECT SCORE	SCORE													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Heaviness.....	3	..	..	..	..	..	..	..	..	..	..	..	..	..	..
2. Slant.....	5	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Uniformity															
Mixed															
3. Size.....	7	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Uniformity															
Too large															
Too small															
4. Alignment.....	8	..	..	..	..	..	..	..	..	..	..	..	..	..	..
5. Spacing of lines.....	9	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Uniformity															
Too close															
Too far apart															
6. Spacing of words	11	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Uniformity															
Too close.....															
Too far apart															
7. Spacing of letters.....	18	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Uniformity															
Too close															
Too far apart															
8. Neatness.....	13	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Blotches															
Carelessness															
9. Formation of letters...	(26)														
General form.....	8	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Smoothness.....	6	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Letters not closed...	5	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Parts omitted...	5	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Parts added.....	2	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Total Score.....	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..

FIG. 9. STANDARD SCORE CARD FOR MEASURING HANDWRITING
 (Devised by C. T. Gray)

reading be taken as one factor for the index. Witham's index is given as follows:

$$\text{Quality} = \left(\frac{\text{number of letters written}}{\text{reading time of the examiner}} \right) \text{ plus or minus a correction.}$$

Witham secured specimens by directing the pupils to copy certain material. He asserts that the correlation between the rate of copying and the rate of writing from memory is high. He says further that the scores secured by his formula are much more accurate and reliable than those of any other method used. His judgment of the results is that the Thorndike Scale gave the poorest results, the Ayres Scale a little better, and his plan the most consistent results. His evidence is not convincing.¹

Sackett used as his index of legibility:

$$\frac{\text{Number of syllables read}}{\text{Time}}$$

He found this index of legibility gave less reliable results than the scales.

Gilchrist² offers another index which is secured by dividing 90 by the writing time in seconds plus nine times the reading time in seconds. Thus, if writing time is 53 seconds, reading time 6 seconds, the formula would be:

$$\text{Index} = \frac{90}{53 + (9 \times 6)} = .84.$$

Methods of using handwriting scales. The quality of a sample of a pupil's handwriting is measured by comparing it with the specimens which compose a handwriting scale. The measure of the quality of the sample is the scale value

¹ Witham used mean deviations for making his comparisons. See Kelley's criticism of this method. Kelley, T. L., "Comparable Measures"; in *Journal of Educational Psychology*, vol. 5, pp. 589-95. (December, 1914.)

² Gilchrist, E. P., "A Handwriting Scale"; in *School and Society*, vol. 12, pp. 203-04. (September 11, 1920.)

of the specimen of the scale which it most nearly resembles. Several methods of comparing samples with a scale are in vogue. When a teacher is working independently the best procedure is the "sorting method" described by Ayres as follows:

To score samples slide each specimen along the scale until a writing of the same quality is found. The number at the top of the scale above this shows the value of the writing being measured. Disregard differences in *style*, but try to find on the scale the *quality* corresponding with that of the sample being scored. With practice the scorer will develop the ability to recognize qualities more rapidly and with increasing accuracy. If the scoring is done twice, the results will be considerably more accurate than if done only once. The procedure may be as follows: Score samples and distribute them in piles, the 20's in one pile, all the 30's in another and so on. Mark these values on the backs of the papers, then shuffle the samples and score them a second time. Finally make careful decisions to overcome any disagreements in the two scorings.

Another good method, which requires less time, but does not secure as good results as the "sorting method," is the "ascending-descending method." This requires that the sample being examined be moved from the lowest step on the scale toward the higher steps until a scale specimen is reached which the judge decides is superior to the sample being measured. The value of this scale specimen is noted. Then, beginning at the upper end of the scale, the sample is compared again with the steps of the scale until the judge decides the specimen on the scale to be inferior to the sample. The sample then receives the rating represented by the point midway between the step of the scale reached in the ascending, and the step reached in the descending, series of comparisons. For example, working upwards on the Ayres Scale the judge may stop at 70 and working downwards at 50. The specimen in hand would then be rated 60.

This method may be varied by rating all of the samples obtained from a class by working up the scale, recording the judgments on the back of the samples. Next, rate all the samples a second time working down the scale, and record the judgments on the face. Finally, the judge goes over the samples a third time and records the average of the two judgments as the true rating.

A method which will require more time, but one which will secure more accurate results than the methods described above, is one in which a group of three or more persons score the samples independently, using the sorting method. Then the scores assigned by all of the judges to a sample are averaged and the result taken as the true score for that sample. The accuracy of the resulting scores will increase with the size of the groups of judges. All of these methods may be used with either the Ayres or Thorndike scales, and with appropriate modifications with the other scales.

The results of a number of investigations have shown that careful training in a relatively poor method of using a scale will produce a marked improvement in the accuracy of the scores. It should follow that careful training in the use of the sorting and group methods would produce highly accurate results.

Measurement for diagnosis. In using Gray's Score Card and the Freeman Scale, separate measures of the factors concerned in a pupil's handwriting are secured. Such diagnostic measures show just which abilities have not been sufficiently improved. These abilities will then be the points of attack for the teacher and pupil in their subsequent work. For example, a record as shown on the Gray Score Card might indicate that a pupil's handwriting was suffering chiefly because of poor letter formation. A closer inspection would show that letter formation was very often defective in two items, letters not closed and parts omitted. Such diagnosis reveals a definite problem for the teacher.

Use of the score card. The score card (see page 173) may be used for a pupil, or a class. If it is used for a pupil, the numerals along the top may be taken to indicate weeks, months, or other intervals. In the column under the numeral 1 the first scores of a pupil's handwriting should be entered. A month later a second series of scores should be entered in the column headed by the numeral 2. The next month another series of scores should be entered under numeral 3, and so on. At the close of a term there will appear a very useful record of the child's experience in the learning of handwriting. This use of the score card Gray calls a "clinical study."

If the card is used for a class, the numerals at the head of the columns stand for the pupils of the class. The totals at the bottom will furnish an interesting comparison of the abilities of the pupils. Each pupil knowing his number can tell how he stands in relation to the other members of the class. If a new score card is posted each month, a pupil may see whether he is gaining or losing in his position in the class. If he is losing, he will be inclined to seek the reason. He may see that his neatness has a low score. This furnishes a strong incentive for work to improve in neatness. Teachers and supervisors might compare their records. The use of the card may be varied by training pupils to score their own or others' handwriting, or by one teacher calling on another teacher to score the handwriting of her pupils.

The individual record card shown in Fig. 10 is a very simple form of a score card designed to be used with the Freeman Scale.

Using the Freeman Handwriting Scale. This scale may be used for measuring the handwriting of all members of a class, but frequently it is desirable to use it only to diagnose the samples written by those ranking conspicuously below the grade norms. This needy group of pupils may be se-

Pupil's Name.....City.....

	First trial Date....	Second trial Date.....	Third trial Date....	Fourth trial Date.....	Building.....	Grade.....	Teacher.....	Age.....
Chart I (Slant)								
Chart II (Alignment)								
Chart III (Quality of line)								
Chart IV (Letter formation)								
Chart V (Spacing)								
Total (value on Freeman Scale)								
Quality (value on Ayres Scale)								
Rate (Letters per minute)								

FIG. 10. INDIVIDUAL RECORD CARD, FREEMAN SCALE

lected by the teacher's unaided judgment, but preferably by the use of the Thorndike or Ayres scales.

Freeman ¹ has issued the following suggestion for using his scale.

The specimen to be judged is graded according to each category separately and given the rank of the specimen in the chart with which it most nearly corresponds in each case. The total rank is calculated by summing up the five individual ranks. Thus, if letter formation is given double value, the lowest possible rank is 6 and the highest possible rank is 30 ($5 + 5 + 5 + 10 + 5$), and the range is 24.

Several precautions are to be observed in making the judgments. The value of the method rests upon the fact that different features of the writing are singled out, one at a time, and graded by being

¹ Freeman, F. N., *Experimental Education*, p. 86. (Boston: Houghton Mifflin Company, 1916.)

given a rank in one of only three steps. The differences between the steps are marked, and the ease of placing a specimen should be correspondingly easy.

This method implies, however, that

- (1) The attention is fixed on only one characteristic at a time.
- (2) The judgment on one point be not allowed to influence the judgment on the other point.
- (3) The same fault be counted only once.
- (4) General impressions be disregarded.

The scores secured by means of the Freeman Scale should be saved to furnish a means of evaluating the results secured from instruction. The scores may be recorded on the specimen, or better on an individual record card, such as shown in Fig. 10. The latter will be more convenient when the teacher wishes to examine a series of scores recorded at intervals over a term of several months.

V. THE RELIABILITY OF MEASURES OF HANDWRITING

The reliability of measures of handwriting is highly important. In the absence of scales such as described here teachers' judgments of the degree of achievement in handwriting have been highly subjective. Handwriting scales have been announced as instruments for making objective measurements. Even limited experience with the scales will convince one who is at all critical that the measures secured are far from perfectly objective. However, handwriting scales are justified if they help teachers to make better estimates of quality than they would make by their unaided judgments, but in order to use the measures intelligently, one needs to be informed concerning the degree of reliability which may be expected.

Reliability of measurements of rate. The rate of a pupil's writing in terms of letters per minute is a definite measure. If no error was made in timing, the rate score tells accurately the average rate at which the pupil was writing

when the sample was secured. It is a true measure of the pupil's normal rate of writing unless the pupil wrote more slowly or more rapidly than he is accustomed to write. Hence, in the measurement of rate of writing we have to consider the question, How nearly were the pupils writing at their ordinary rate?

It is certain that the conditions under which the pupils write influence their rate very materially. As quoted above (page 163), Sackett found that when pupils know they are being tested the rate of writing is reduced. Freeman's results and the results secured by the writer show that the character of the instructions given the pupils is very potent in determining their rate of writing. For example, in response to the instruction, "Write as fast as you can," one college sophomore increased her rate of writing seventy-seven letters a minute over her rate when writing for highest quality. If such conditions as the knowledge that one is being tested and instructions to "write rapidly" or "write well" influenced all pupils to the same degree, we could disregard them in considering the reliability of measures of rate. This, however, does not occur. The truth of this statement may be tested by securing three samples of handwriting from a class: the first written in response to the direction, "Write as well as you can"; the second written in response to the direction, "Write as rapidly as you can"; and the third, ordinary writing. Three such samples were secured from eighty-eight pupils representing the fourth, sixth, and eighth grades. The rate of their ordinary writing compared with the rate of writing at a maximum speed gave a correlation of .586. The rate of writing for their best quality compared with their writing at greatest speed gave a correlation of .5155. The rate of writing for quality compared with the rate of ordinary writing gave a correlation coefficient of .724. The P.E.'s of these coefficients are so

low as to be insignificant. Hence, writing at highest speed is an ability or a combination of abilities more different from the abilities concerned in writing for quality or the abilities involved in what we have called ordinary writing than these later abilities are different from each other. It may be that these abilities are analogous to the abilities to be a graceful dancer, a rapid walker, or a fast runner. The three operations just named all involve similar neuro-muscular patterns, but are not closely associated in their varying degrees of excellence. If this suspected condition should prove to be the true one, it will demand that we train children in the type of writing which they will most often have occasion to use.

Reliability of measurement of quality: scales compared. A number of investigations have been made to determine the effect of the particular scale used upon the reliability of the scores assigned to samples of handwriting. Most of these studies have compared the Thorndike and Ayres scales with each other or with other measuring devices. Pintner¹ reported the Thorndike Handwriting Scale as being superior to the Ayres Three Slant Edition. Kelley,² using Pintner's data and making use of a refinement of method, showed that the uniformity of estimate of handwriting by means of the Ayres Scale is slightly greater than by means of the Thorndike Scale. For the average of two sets of ratings by several judges, using the two scales, a coefficient of correlation of .98 is reported by Pintner. This is a very high agreement. In comparing the Ayres Three Slant Edition with the Gettysburg Edition, Breed³ found the latter scale gave

¹ Pintner, R., "A Comparison of the Ayres and Thorndike Handwriting Scales"; in *Journal of Educational Psychology*, vol. 5, pp. 525-31. (November, 1914.)

² Kelley, T. L., "Comparable Measures"; in *Journal of Educational Psychology*, vol. 5, pp. 589-95. (December, 1914.)

³ Breed, F. S., "The Comparable Accuracy of the Ayres Handwriting Scale, Gettysburg Edition"; in *Elementary School Journal*, vol. 18, p. 458. (February, 1918.)

more reliable ratings. Sackett¹ found the Ayres Adult Scale to be most reliable, the Three Slant Edition second in reliability, and the Thorndike Scale third. As the samples which he used were written by university sophomores the Ayres Adult Scale was doubtless best adapted to that kind of handwriting. From all these studies we may conclude that Ayres Handwriting Scale, Gettysburg Edition, may be expected to yield slightly more accurate measures of quality than either the Thorndike Scale or the other scales by Ayres. However, it should be recognized that we have available several scales for the measurement of handwriting which are about equally reliable when used under similar conditions.

Objectivity of scores assigned by judges. Reliable measurements of the quality of handwriting depend more on the scorer than on the particular scale used. The teacher must keep in mind that the issue here is not "perfect measurement or none," but rather, "How can we make our measurements more reliable?" There is abundant evidence² that untrained persons differ considerably in the scores assigned to the same samples of handwriting. On the other hand, it is generally agreed that a teacher's estimate of the quality of a sample of handwriting, made with the aid of a scale, will be more accurate than her unaided judgment. Morton³ studied 31,100 ratings of samples of handwriting. The use of

¹ Sackett, L. W., "Comparable Measures of Handwriting"; in *School and Society*, vol. 4, pp. 640-45. (October 21, 1916.)

² Breed, F. S., and Culp, Vernon, "An Application and Critique of the Ayres Handwriting Scale"; in *School and Society*, vol. 2, pp. 36-47.

Manual, H. T., "The Use of an Objective Scale for Grading Handwriting"; in *Elementary School Journal*, vol. 15, p. 269. (January, 1915.)

King, Irving, and Johnson, Harry, "The Writing Abilities of the Elementary and Grammar School Pupils of a City School System Measured by the Ayres Scale"; in *Journal of Educational Psychology*, vol. 3, pp. 514-20. (October, 1912.)

³ Morton, R. L., "The Value of a Handwriting Scale to an Untrained Teacher"; in *Journal of Educational Research*, vol. 3, p. 133. (February, 1921.)

the scale by teachers untrained in its use reduced their average error five per cent below that of their unaided judgment. Kelly,¹ Lewis,² and Gray³ agree that with equal training in the use of the usual percentile method of grading, and in the use of a scale, the results of the latter method are more accurate.

Training in the use of a handwriting scale. Thorndike⁴ has published fifty samples of handwriting whose quality has been determined in terms of his scale. They are to be used in training teachers in the use of a handwriting scale. Thorndike's plan is for a teacher to score each sample without referring to its true value. He then compares his score with the true value and notes his error. This is done with each of the fifty samples. Then the samples are to be taken up in random order and all scored again as before. Each time they are scored there will be some gain in the accuracy of the scores. Hurt⁵ found that three weeks of such training enabled seventeen of twenty-one judges to reduce their errors within negligible limits.

Hurt also carefully tested the effects of several methods of training in the use of the Thorndike Scale. He has shown that independent practice in the use of the scale reduces the individual variation of several ratings of one set of samples. It was also shown that after two groups of judges had had

¹ Kelly, F. J., *Teachers' Marks*. Teachers College Contributions to Education, no. 66, pp. 99-108. (New York: Teachers College, 1914.)

² Lewis, E. E., "The Present Standard of Handwriting in Iowa Normal Training High Schools"; in *School Administration and Supervision*, vol. 1, pp. 663-71. (December, 1915.)

³ Gray, C. T., "The Training of Judgment in the Use of the Ayres Scale for Handwriting"; in *Journal of Educational Psychology*, vol. 6, pp. 85-98. (February, 1915.)

⁴ Thorndike, E. L., "Teachers' Estimates of Specimens of Handwriting," *Teachers College Record*, vol. 15, no. 5. (November, 1914.)

⁵ Hurt, A. O., *A Study of the Reliability of the Thorndike Handwriting Scale*. (An unpublished Master's Thesis, University of Missouri.)

several weeks of practice, the group which practiced two weeks longer succeeded in still further reducing their individual variation or variations with their own previous ratings. It seems probable that an individual can make his rating more consistent by a long period of independent practice. But such consistency may not make one judge's ratings agree any better with the ratings of other judges if other judges have not been subjected to the same training.

Gray¹ tested the effect of practice with instruction in the use of the Ayres Scale. He gave three judges careful instruction and practice for twenty weeks. At the end of ten weeks the group variation was a little more than half of what it was the first week. The twentieth week found their variation reduced to one seventh of the first week's variation. Gray says:

Accuracy in grading writing by a scale may be produced by careful training in the use of the scale. In the past the assumption has been made that ability to grade expertly in a subject came with an expert knowledge of the subject. While the experiment does not disprove this assumption, it indicates clearly that another avenue of approach to such expert ability is through a period of careful training. This implies that grading may be considered a field more or less by itself, and gives a glimpse of a type of work in education whose chief interest is *the accurate use of units of measurement*.²

VI. NORMS³

Two types of norms. For handwriting, it is well to recognize *norms of progress* and *norms of attainment*. A norm of

¹ Gray, C. T., "The Training of Judgment in the Use of the Ayres Scale of Handwriting"; in *Journal of Educational Psychology*, vol. 6, pp. 85-90. (February, 1915.)

² Italics ours.

³ No attempt is made to give A.Q.'s or E.Q.'s for handwriting. Scores in handwriting have about zero correlation with intelligence scores, and relatively low correlation with age.

progress is the degree of handwriting ability which should exist at a given grade or age level. A *norm of attainment* is the degree of the ability which should exist when the training period for that ability is completed. Thus, in Table XVIII we have one set of *norms of progress*, or the scores in quality and rate of handwriting which should be attained by the pupils in the grades indicated. These norms imply a certain *norm of attainment* to be reached in the eighth grade.

TABLE XVIII. NORMS OF PROGRESS PROPOSED BY FREEMAN ¹

	SCHOOL GRADES						
	II	III	IV	V	VI	VII	VIII
Quality . . .	44	47	50	55	59	64	70
Rate	36	48	56	65	72	80	90

Norms of progress. Freeman has proposed the norms of progress given in Table XVIII. They were not based on median scores, but were determined by the median of the up-

TABLE XIX. NORMS OF PROGRESS FOR THE AYRES
GETTYSBURG EDITION

	SCHOOL GRADES						
	II	III	IV	V	VI	VII	VIII
Quality . . .	38	42	46	50	54	58	62
Rate	32	44	56	64	70	76	80

¹ Freeman, F. N., *Fourteenth Yearbook of the National Society for the Study of Education*, part I.

per half (75 percentile) of the scores for each grade. About five thousand specimens of handwriting were scored for each grade. These were gathered from fifty-six cities. Freeman also investigated the demands made upon persons employed in several large commercial houses. The results of this investigation were considered in setting up the *norms of progress* given in Table XIX. He estimated that these norms could be attained by an expenditure of not over seventy-five minutes per week. Later Freeman ¹ directed a survey in which a single trained grader scored the samples of handwriting from one school in each of the fifty-six cities, and a larger sampling from three large cities. The median scores found were but slightly different from the medians as given for the fifty-six cities in the earlier study, as shown in Tables XX and XXI.

TABLE XX. MEDIAN SCORES FOUND (RATE)

	SCHOOL GRADES							APPROXIMATE NUMBER OF SPECIMENS SCORED
	II	III	IV	V	VI	VII	VIII	
Cleveland ¹	..	..	..	60	70	76	80	25,387
Iowa schools ²	39.2	49.2	61.9	65.5	72.6	75	76.5	28,000
Starch's Standards ³	31	38	47	57	65	75	83	4,740
Kansas Medians ⁴	32	35	51	61	67	71	73	6,000
Fifty-six cities ⁵	30.6	43.8	51.2	59.1	62.8	67.9	73	34,000
Freeman's Standards.....	36	48	56	65	72	80	90	

¹ Judd, Charles H., *Measuring the Work of the Public Schools*. (Report, Survey Committee of the Cleveland Foundation, 1916.)

² Ashbaugh, E. J., *Handwriting of Iowa School Children*. (University of Iowa, Extension Division, Bulletin no. 15, March, 1916.)

³ Starch, D., *The Measurement of Efficiency in Reading, Writing, Spelling, and English*. (University of Wisconsin, 1914.)

⁴ DeVoss, J. C., *Second Annual Report of Bureau of Educational Measurements and Standards*. (Kansas State Normal School, Emporia, Kansas.)

⁵ Freeman, F. N., *Fourteenth Yearbook of the National Society for the Study of Education*, part I. (1915.)

¹ Freeman, F. N., *Sixteenth Yearbook of the National Society for the Study of Education*, part I.

TABLE XXI. MEDIAN SCORES FOUND (QUALITY)

	SCHOOL GRADES								APPROXIMATE NUMBER OF SPECIMENS SCORED
	II	III	IV	V	VI	VII	VIII	Scale used	
Cleveland ¹				45	48	50	55	Ayres	25,387
Iowa ²	35.7	39.8	44.5	49.1	52.3	57	61	"	28,000
Starch's Standards ³	27	33	37	43	47	53	57	"	4,740
Kansas Medians ⁴	44	47	50	55	59	64	70	"	6,000
Fifty-six cities ⁵	39.7	42	45.8	50.5	54.5	58.9	62.8	"	34,000
Freeman's Standards (Ayres' Scale).....	44	47	50	55	59	64	70	"	
Salt Lake City ⁶		9.2	10.7	11.1	11.3	12.2	12.8	Thorn- dike	2,500
Butte, Montana ⁷	8.2	8	8.8	8.9	11.6	11.2	12.1	"	1,400
Southington, Conn. ⁸						10		"	
Connersville, Ind. ⁹		10.3	10	10.3	11.7	11.7	11	"	1,200
Freeman's Standards (Thorndike's Scale)	9.36	9.75	10.13	10.76	11.34	11.89	12.66		

¹ Judd, Charles H., *Measuring the Work of the Public Schools*. (Report, Survey Committee of the Cleveland Foundation, 1916.)

² Ashbaugh, E. J., *Handwriting of Iowa School Children*. (University of Iowa, Extension Division, Bulletin no. 15, March, 1916.)

³ Starch, D., *The Measurement of Efficiency in Reading, Writing, Spelling, and English*. (University of Wisconsin, 1914.)

⁴ DeVoss, J. C., *Second Annual Report of Bureau of Educational Measurements and Standards*. (Kansas State Normal School, Emporia, Kansas.)

⁵ Freeman, F. N., *Fourteenth Yearbook of the National Society for the Study of Education*, part 1. (1915.) Revised medians are given in the *Sixteenth Yearbook of the National Society for the Study of Education*, part 1. (1916.)

⁶ *Report of a Survey of the Schools of Salt Lake City, Utah*. (1915.)

⁷ *Report of a Survey of the Schools of Butte, Mont.*, chap. iv. (1914.)

⁸ Witham, E. C., "All the Elements of Handwriting Measured"; in *Educational Administration and Supervision*, vol. 1, pp. 313-24. (May, 1915.)

⁹ Wilson, G. M., "The Handwriting of School Children"; in *Elementary School Teacher*, vol. 6, pp. 540-43. (1911.)

Other evidence as to norms of progress. Tables XX and XXI give the results of a number of widely scattered investigations, and show the median scores found in these different places. The Freeman norms are inserted in each table for comparison. The figures in the columns at the extreme right show the total number of samples rated in each investigation. A comparison of the median scores shown in these tables with the norms proposed by Freeman (Table XIX), shows that the latter are higher in most instances. This, together with other evidence, points toward a possible modification of the norms of progress set by Freeman.

Norms of attainment. Ayres¹ and Ashbaugh² have drawn certain conclusions from the requirement in handwriting which are set by the examiners of the Municipal Civil Service Commission of New York City. Ashbaugh quotes a letter from the Acting Director of the commission as follows:

I find that the Municipal Civil Service Commission of New York ordinarily uses the standard of 70 per cent as a passing grade in handwriting, but for positions where handwriting is a special requirement the standard is sometimes set at 75 per cent.

Ayres has shown that the ratings of 70 per cent and 75 per cent, as given by the commission, correspond respectively to scores of 40 or 50 on the Ayres Scale. Since this commission recommends many persons who cannot write better than the 40 specimen of the Ayres Scale, and recommends others who write only as well as the 50 specimen for positions where handwriting is a special requirement, it would follow that an ability to write as well as 50 on the Ayres Scale would be sufficient for all the demands which many pupils will meet.

Lewis³ examined the handwriting of 1760 third- and fourth-year students of 166 Iowa Normal Training high schools. He found their median score for quality to be 59.1 on the Ayres Scale, with a range from 34 to 89. Fifty per cent of the scores fell between 53.6 and 64.3. The average rate of their handwriting was ninety letters per minute. Comparing these median scores with those found by Ash-

¹ Ayres, L. P., *A Scale for Measuring the Quality of Handwriting of Adults*. Russell Sage Foundation Bulletin no. 138. (New York: Russell Sage Foundation.)

² Ashbaugh, Ernest J., *Handwriting of Iowa School Children*. University of Iowa Bulletin no. 110. (Iowa City: University of Iowa, 1916.)

³ Lewis, E. E., "The Present Standard of Handwriting in Iowa Normal Training High School"; in *Educational Administration and Supervision*, vol. 1, pp. 663-71. (December, 1915.)

baugh (Tables XX and XXI) in Iowa elementary schools, we may ask, "Did eighth-graders, with a median rate of 76.5 letters per minute and with a quality of 61, lose some in their quality in order to raise their rate to ninety letters per minute in high schools?"

Koos¹ investigated 1053 specimens of handwriting taken from social correspondence and 1127 specimens written by employees in various occupations. He also secured the judgment of 826 adults as to what they considered adequate and inadequate handwriting for social correspondence. He concluded from the study of social correspondence that "it is difficult to see why, for the use under consideration, a pupil should be required to spend the time necessary to learn to write better than the quality 60. There is even considerable justification for setting the ultimate standard at 50. As this demand touches every member of society, all the children in the schools should be required to attain the standard set."

From his study of vocational uses of handwriting, Koos concludes that "the quality 60 on the Ayres Measuring Scale for Adult Handwriting, which we have set up as the ultimate standard of attainment for all school children for purely social purposes, is adequate for the needs of most vocations. This will apply to labor, skilled and unskilled, as well as to the professions, exclusive of teaching in the elementary schools. For that large group who will go into commercial work, for telegraphers, and for teachers in the elementary schools, it will be necessary to insist upon the attainment of a somewhat higher quality, but hardly in excess of the quality 70." Thus the quality of 60 appears to be sufficient for the basic needs which are to be satisfied by the general education of the elementary schools. Those pupils

¹ Koos, L. V., "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools"; in *Elementary School Journal*, vol. 18, p. 422. (February, 1918.)

who require the quality of 70 should receive this through vocational training in the schools which fit them for their specific fields of work.

Almack ¹ found that the average quality of handwriting of 1711 rural teachers of Oregon was 12.75 on the Thorndike Scale, which would be equivalent to 70 on the Ayres Scale.

Nifenecker ² found 106 letters per minute to be the average rate for 161 sales clerks, billers, checkers, and bookkeepers. This represents a vocational demand for rate of writing and should be compared with the rate secured when the pupils are asked to write rapidly. Like quality 70, it should be secured through vocational training.

The study by Lewis suggests ninety letters per minute as a "norm of attainment" for those who expect to go on to high school. High school and college demand a higher rate of handwriting than is often demanded in the lower grades. Teachers of penmanship generally agree that quality should be the first goal. Having attained the standard quality, the norm of ninety letters per minute may be considered a legitimate goal.

Norms of attainment for maximum rate and maximum quality. A valuable measure of handwriting ability can be secured by calling for two specimens, one written at maximum rate at which legible handwriting can be produced, and one written for maximum quality. If the maximum of rate is ninety letters per minute or over, and the writing is sufficiently legible for purposes of note-taking, the pupil's ability is probably adequate for high-school demands. If the maximum quality specimen rates sixty or over on the Ayres Scale, and is written at a rate of seventy or more letters per minute, the pupil's skill in handwriting is probably adequate for all social and vocational demands.

¹ Almack, John C., "Writing Ability of Teachers"; in *School and Society*, vol. 10, pp. 389-90. (September 27, 1919.)

² *Op. cit.*

VII. PROCEDURE AFTER MEASUREMENT

Extremes to be avoided in the interpretation of scores. The usefulness of educational measurements depends largely on the way in which the scores are treated. Two extremes must be avoided here. The one of careless neglect of any arrangement of scores will certainly lead to a condemnation of measurement as a fad. There are doubtless many unused scores lying about in the files of school principals' offices and

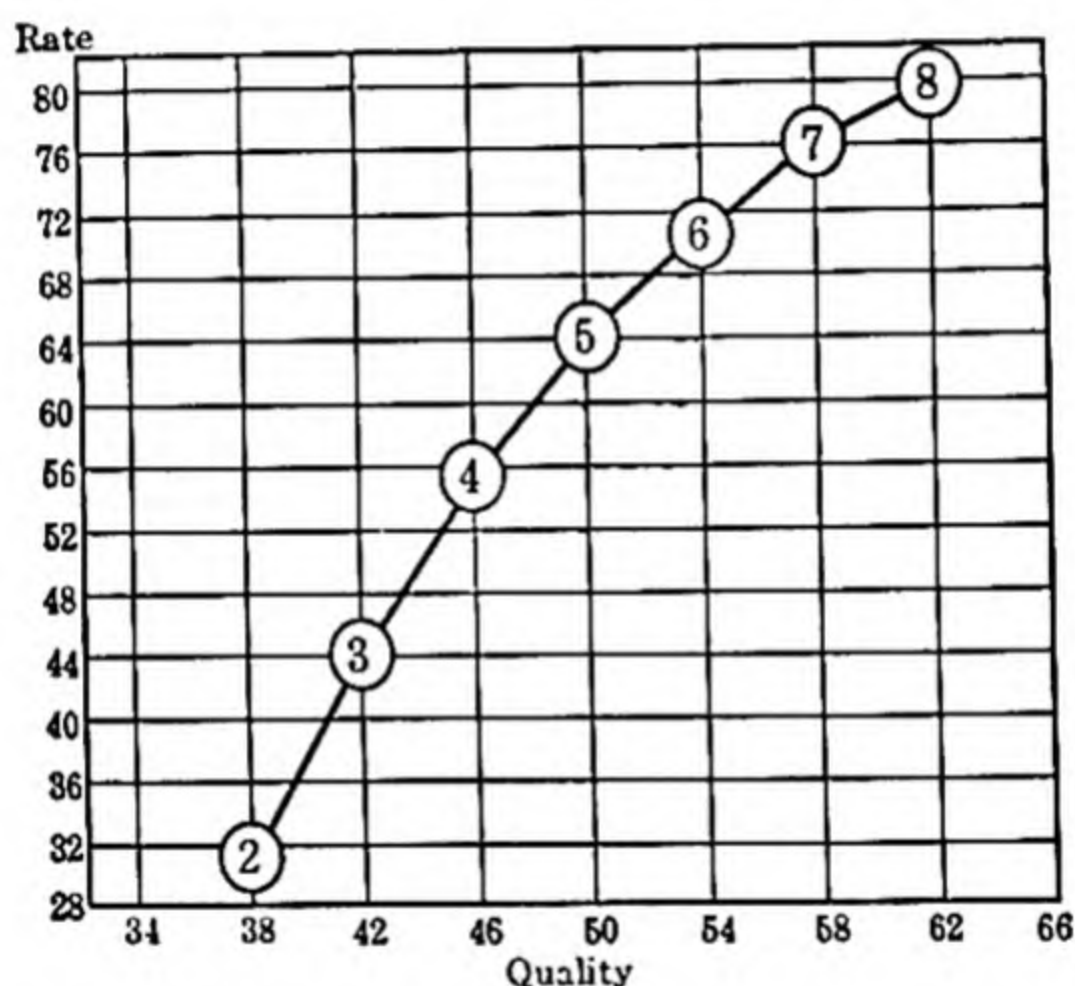


FIG. 11. GRAPHICAL REPRESENTATION OF GRADE NORMS GIVEN BY AYRES FOR THE GETTYSBURG EDITION OF HIS SCALE

of bureaus of research, which represent almost a dead loss of time and energy. The other extreme is that of elaborate charts and elaborate statistical treatment which render little or no assistance in the interpretation of the scores. Such treatment of scores may have its place in the work of the highly trained investigator, but it is not necessary for the interpretations which the teacher should make, and no teacher

should be discouraged by the failure to understand the elaborate treatments of scores found in some of our literature.

Interpreting the scores of a class. At this point let us assume that samples of handwriting have been secured from a sixth-grade class under standard conditions, including the careful use of detailed instructions as described earlier in this chapter, and that the scores of rate in letters per minute and of quality on the Ayres Gettysburg Scale have been recorded in one corner of each sample. How can we best display these scores for further study? First let us stack the samples in the order of the quality scores. In this class we find two pupils received a quality score of 80, four 70, etc., as below:

SCORES	NUMBER OF PUPILS
80	2
70	4
60	7
50	7
40	7
30	4
20	2
—	—
Total 33	

We shall put all the papers with the quality score of 80 in one group, then those of 70 in another, and so on. Next we shall pile them as shown above with the two scoring 20 at the bottom, and the two scoring 80 at the top. Finding the middle paper, we see that it has a score of 50, so we know the median and average scores will be not much larger or smaller than 50. Referring to the norms given on the Ayres Gettysburg Edition, we find the median slightly below the standard proposed there. (See Fig. 11.) But we recall from our sorting of the papers that some were far below 50 and some scored as high as 80. How can we emphasize the location of individual pupils? Let us take some cross-section paper

and lay off along the base line a space for each quality score which has occurred in this class, 20 to 80. Then, allowing a square to each pupil, let us write in the first name and initial of surname for each pupil. We shall then have a histogram as in Fig. 12. Here we have a clear picture of the quality scores of the class. To assist in the interpretation of the individual scores both the class median and the sixth grade median (norm) have been included. The median for the class is slightly below the norm for the grade.

Next the samples should be resorted according to the rate scores, placing in one pile all papers with rate scores from 27.5 to 32.5, in another pile those from 32.6 to 37.5, etc. By following a procedure similar to that described in the above paragraph, we may construct the histogram for rate as shown in Fig. 13.

With these two pictures before us, we can study the individual needs of our pupils. It is obvious that Lynn M. is writing too fast for the quality he scores. We can try Lynn M. again, insisting that he write better. On the other hand, Laura S. is writing a superior quality at a rate beyond the standard of her grade. As she is not over age for her grade, we can feel sure that, with very little guidance, Laura S. will make all the progress necessary in the seventh and eighth grades.

If we wish to make the picture a little more rapidly, we may make a chart like Fig. 14. This "two-way chart" is usually a little more difficult for pupils to read; hence, if a part of our purpose is to show the pupils their standing, the histograms will be more useful. But the "two-way chart" has this advantage that it shows at a glance that those above the heavy line representing rate 70 (sixth-grade norm) and to the right of quality 54 (sixth-grade norm) are doing very satisfactory handwriting. In this class only two are in that part of the distribution. Those on the lower and right-hand side

of the standard lines are writing too slowly, but with sufficient quality. Other specimens should be secured from them by asking them to write more rapidly in order that we may discover whether or not they can write a satisfactory quality at an increased rate.

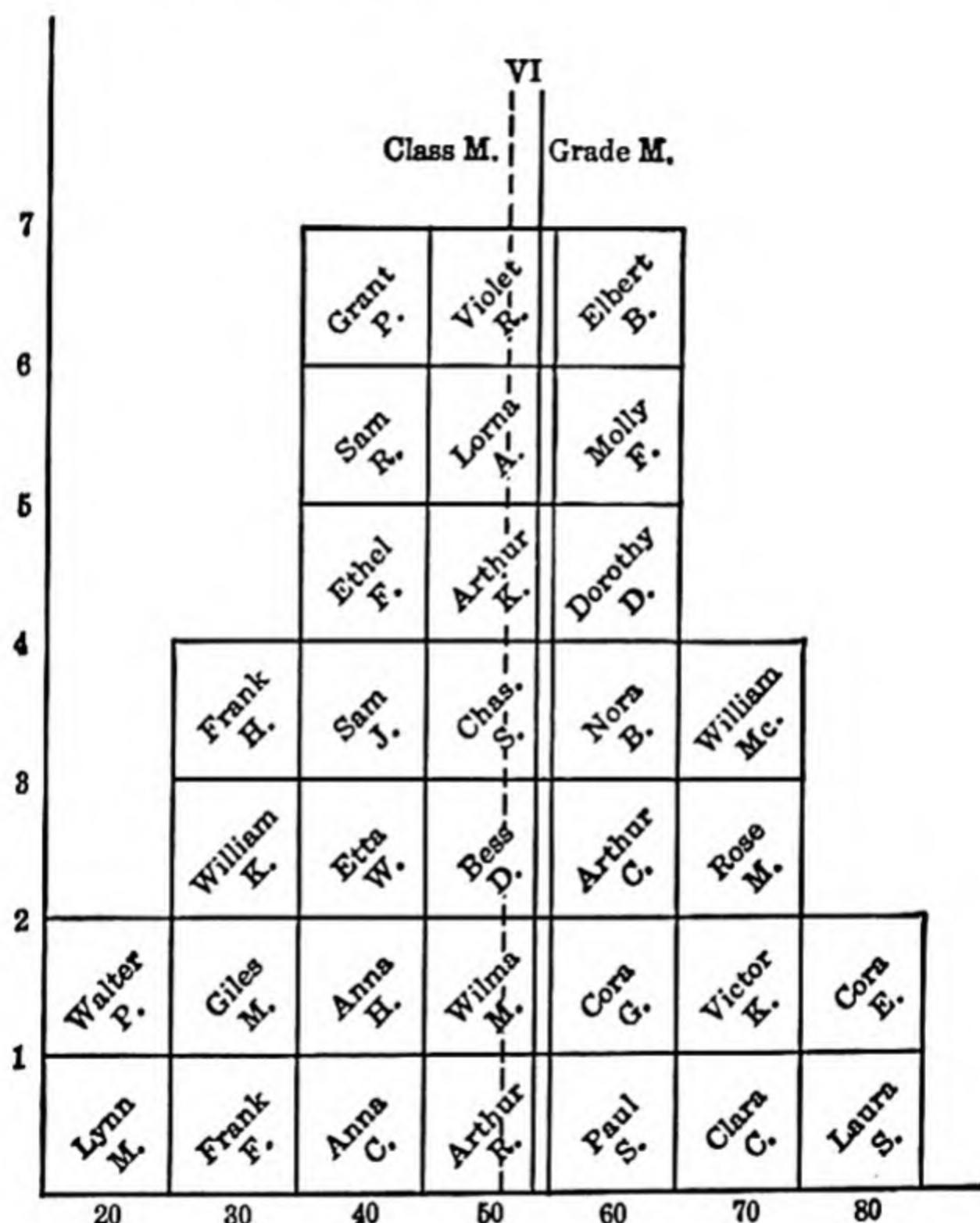
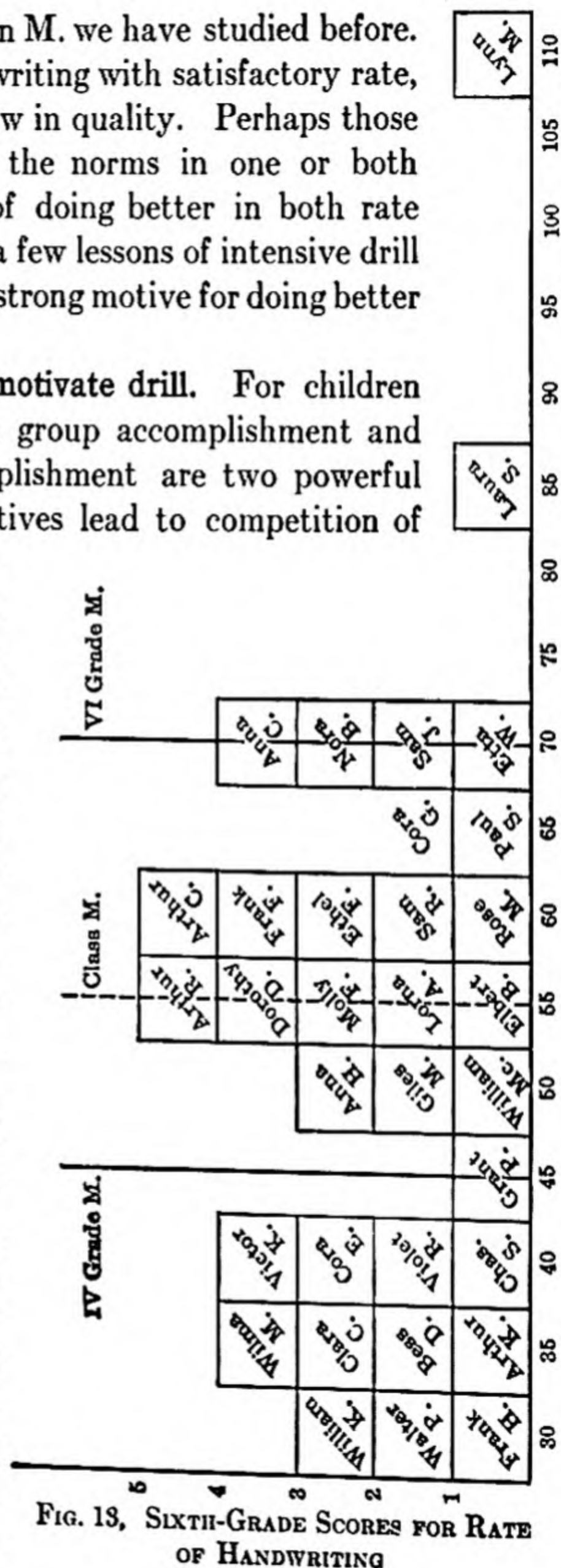


FIG 12. SIXTH-GRADE SCORES FOR QUALITY OF HANDWRITING (AYRES SCALE)

Those in the lower left-hand section should be writing more rapidly and with better quality unless some of them are bright children who are younger than the average for the grade. Those in the upper left-hand section are obviously

in two groups. Lynn M. we have studied before. The other three are writing with satisfactory rate, but are just a step low in quality. Perhaps those who are just below the norms in one or both scores are capable of doing better in both rate and quality if given a few lessons of intensive drill and furnished with a strong motive for doing better work.

Using scores to motivate drill. For children of this age pride in group accomplishment and pride in self-accomplishment are two powerful motives. These motives lead to competition of group against group and to competition against one's own record. There are several good ways to stimulate group competition. The median scores of the class may be compared with the median scores of other classes and cities. For this purpose Tables XX and XXI are useful. The norms given in Fig. 11 and in Table XIX are also useful. Our sixth grade may see that in quality their score is as good as that of the Cleveland sixth grades,



but not so good as the medians of the fifty-six cities. To lead the class to see that all are included in this game, show them how their own distribution compares with the standard distribution displayed at the bottom of the Gettysburg

Rate	Quality			VI Grade M.				Total	
	20	30	40	50	60	70	80		
101-110	Lynn M.							1	
91-100								0	
81-90							Laura S.	1	
71-80			Etta W. Sam J. Arthur C.		Nora B.			4	VI Grade M.
61-70		Frank F.	Sam R. Ethel F. Anna C.		Paul S. Cora G.	Rose M.		7	
51-60		Giles M.	Anna H.	Lorna A.	Elbert B. Molly F. Dorothy D.	William Mc		7	
41-50			Grant P.	Arthur R. Violet R. Chas. S.		Victor K.	Cora E.	6	
31-40	Walter P.	Frank H. William K.		Wilma M. Beas D. Allen K.		Clara C.		7	
Total	2	4	8	7	6	4	2	33	

FIG. 14. SIXTH GRADE SCORES FOR RATE AND QUALITY OF HANDWRITING

Scale. Such comparisons carried out at least once a month will stimulate unusual interest in any subject which is measured.

Self measurement. Many teachers have reported good results from various plans in which each pupil would compete with his own past record. Henmon reports a study in-

cluding several of the school subjects. He concludes that much of the gain reported was due to the incentive which resulted from the pupils knowing their progress. Henmon¹ says:

The effect of applying the tests at regular intervals throughout the year is nothing short of remarkable.

The results are in accord with those attained in laboratories which have always shown very great improvement with practice under experimental conditions even in traits which in the ordinary circumstances of life have been much subject to practice. Laboratory experiments in memorizing, for instance, have shown very great gains under experimental conditions by those who have memorized more or less all their lives and whose capacities might be supposed to have been brought near the limits of improvement.

Certain practical applications of these results suggest themselves. Would it not be desirable to place a great deal of school work in the form of a practice experiment in which the pupil tested himself at regular intervals by objective methods, kept his own score, and watched his gain or loss from week to week or month to month? All experiments show that this would be an extraordinary stimulus to effort.

The display of handwriting scales on the walls of the room has been advocated and practiced with good results. Freeman² has prepared handwriting and measuring tablets which display a limited scale in the practice tablet. Freeman says:

The purpose of this scale and of its inclusion in the pupil's writing book is to make it convenient to require the pupil frequently to measure his own writing. The presence of the scale in his blank book, where he looks at it and uses it frequently, makes the child so familiar with the specimens which represent different

¹ Henmon, V. A. C., "Improvement in School Subjects throughout the School Year"; in *Journal of Educational Research*, vol. 1, pp. 81-95. (February, 1920.)

² Freeman, F. N., "A Handwriting Scale for the Pupil"; in *Elementary School Journal*, vol. 21, pp. 755-61. (June, 1921.) Supplied by Dodson Evans Company, Columbus, Ohio.

grades of writing, that a given score means something to him besides an abstract number. It stands for a concrete representation of the quality in question.

It is good practice to supplement the rating of the pupil's writing which is made in terms of a scale, by keeping a series of specimens written at monthly intervals, so that he may discern progress which is finer than can be expressed in units of the scale.

This type of concrete progress record also enables the pupil to analyze his improvement and to trace the particular aspects of his writing in which he has gained. Especial emphasis should be placed in teaching on the pursuit by the pupil of special aims, such as the increase in uniformity of slant or alignment, the improvement of certain letter forms or of the spacing between words or lines. Such definiteness gives meaning to practice and makes progress more perceptible.

The Courtis Standard Practice Tests in Handwriting¹ furnish a carefully prepared plan for motivating practicing in handwriting through group competition and competition with one's own records. These tests make use of self-measurement in a very effective way.

Directing the practice drills. The measurement of the rate and quality of pupils' handwriting naturally leads to the situation in which those who fall below the norms of progress will ask how they can learn to write better. It is beyond the province of this chapter to describe methods of teaching handwriting. There are many excellent manuals of instruction. These manuals and any other aids or plans for practice drills should be criticized and selected or rejected in the light of facts which have been established by investigations of the learning process as it occurs in learning to write.

Systems of penmanship. There are not sufficient data from comparative studies of different penmanship systems to establish any single system as superior to others in its ef-

¹ Courtis, S. A., *Standard Practice Tests in Handwriting*. (Yonkers: World Book Company.)

fectiveness to secure results in terms of speed and quality of handwriting.

Movement in handwriting. Graves¹ describes three kinds of movement: first, finger-movement; second, "arm-movement" in which there is some movement of the fingers and considerable movement of the arms; and, third, free-arm-movement, in which "the respective movements of the fingers and of the arm are proportionally equal in amount." Of these three types of movement Graves concludes that arm-movement seems to give greater speed. Nutt² does not find this to be so. Since Nutt secured a positive result of no correlation between speed and movement, and since his measuring devices and methods were more objective than those used by Graves, it seems safe to conclude that movement does not influence speed in writing for a short time. The apparent greater ease of production of arm or muscular movement may result in greater speed if speed is measured during a long period of writing.

Nutt also found that arm-movement comes with age and motor development. None of the systems of penmanship were found to develop any appreciable amount of arm movement in younger children. Copy-book methods and the teacher's emphasis on arm-movement develop about the same degree of arm-movement in ages ten to fourteen. Special supervisors secure more arm-movement in children of these ages, and also in nine-year-olds. Well-developed arm-movement did not produce better quality than movements in which the arm was moved but little. Neither did well-developed arm-movement show greater speed. Neither does better arm-movement result in an increase in rhythm.

¹ Graves, S. Monroe, "A Study of Handwriting"; in *Journal of Educational Psychology*, vol. VII, p. 486. (October, 1916.)

² Nutt, H. W., "Rhythm in Handwriting"; in *Elementary School Journal*, vol. 17, pp. 432-45.

The child's natural rhythm of motion is an important factor in his learning to write.

Rhythm. The rhythmic quality of the movement increases with age, but has no relation to amount of arm-movement or to the quality of the writing. Nutt found that speed of writing and rhythm increase together. That is, children who score high in rhythm also score high in speed, are older than the other children, but may not use arm-movement or produce a better quality of handwriting than other children.

Speed. Both Nutt and Graves have shown that speed increases with age. Nutt shows that speed increases with an increase in the rhythmic character of the movement. An important factor in the production of speed of handwriting is that of hand position. Graves says that a free and easy or loose-handed position is most conducive to speed. There is some evidence that girls write more rapidly than boys.

Reasons for using handwriting scales. Even when measures of handwriting are not highly accurate, they direct the attention of the teacher to the specific faults and needs of pupils. Measurement by means of handwriting scales creates in the teacher a more critical and scientific attitude toward the outcomes of instruction. This attitude tends to remove the attention from personal bias and feeling to an objective consideration of the results secured. Measurement of handwriting also banishes the old false standards represented by the perfect specimens which are produced from an engraved plate. In their stead the norms proposed are within the reach of a majority of the pupils. Thus many children can know the joy which comes from achieving something recognized to be of value.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. How can the value of the kind of movement which a pupil uses be scientifically determined?
2. Direct pupils to write as well as they can for two minutes using a good sentence for copy. Next direct them to write as fast as they can using the same copy. Be careful in taking the time. Score these samples for both rate and quality. Determine the largest and smallest differences between the two samples written by a pupil.
3. Study the quality of the samples secured as above.
4. Direct pupils to write one sample writing as well and as rapidly as possible. Do they write as well as they did when told to write as well as they could? Do they write as rapidly as they did when told to write as fast as they could?
5. Collect several long-hand letters written by persons of known social status and success. Rate these samples for quality and compare with report of findings of Koos, Lewis, and Almack.
6. Observe ten cases of persons writing in ordinary situations in stores and offices. What are the demands of such situations?
7. Should pupils use fountain pens in writing exercises? Pencils?
8. Use the Gray Score Card (or Freeman Scale) in scoring specimens of handwriting which are below the grade norms. Prescribe the drills you would use in correcting these defects. Compare this with the recommendations of other teachers or students. Try your prescription on the pupils concerned if possible.
9. For what purpose would you use the dictation exercises?
10. Use the Gray Score Card by filling it with scores secured from use of the Freeman Scale, whenever they apply.
11. Select a defect of letter formation frequently found in a pupil's handwriting. Direct the pupil's attention to this defect and challenge him to correct it. Direct that a record be taken as follows: If the defect were found in letter "a," instruct the pupil to count the number of such errors to be found in fifty consecutive "a's" as they occur in his handwriting written prior to the time you pointed out the defect. After a period of practice, direct the pupil to make another counting from his handwriting written at some period other than the writing period.
12. Collect samples of handwriting from a class and make histograms and two-way chart as shown in Fig. 14.

SELECTED BIBLIOGRAPHY

- Ashbaugh, Ernest J. *Handwriting of Iowa School Children*. University of Iowa Bulletin no. 110. (Iowa City: University of Iowa, 1916. 24 pp.)
- Ayres, L. P. *A Scale for Measuring the Handwriting of School Children*. Russell Sage Foundation Bulletin no. 113. (New York: Russell Sage Foundation, 1912. 16 pp.)
- Ayres, L. P. *A Scale for Measuring the Handwriting of Adults*. Russell Sage Foundation Bulletin no. 138. (New York: Russell Sage Foundation, 1915. 12 pp.)
- Beatty, Willard W. "Judging Handwriting. A Critical Weakness of the Thorndike Scale Revealed by a Comparative Study"; in *Journal of Educational Psychology*, vol. 13, pp. 170-72. (March, 1922.)
- Breed, F. S. "The Comparable Accuracy of the Ayres Handwriting Scale, Gettysburg Edition"; in *Elementary School Journal*, vol. 18, p. 458. (February, 1918.)
- Breed, F. S., and Culp, Vernon. "An Application and Critique of the Ayres Handwriting Scale"; in *School and Society*, vol. 2, pp. 36-47. (October, 1915.)
- Breed, F. S., and Down, E. F. "Measuring and Standardizing Handwriting in a School System"; in *Elementary School Journal*, vol. 17, pp. 470-84. (March, 1917.)
- Freeman, Frank N. *The Teaching of Handwriting*. (Boston: Houghton Mifflin Company, 1915.)
- Freeman, Frank N. "An Analytical Scale for the Judging of Handwriting"; in *Elementary School Journal*, vol. 15, p. 432. (April, 1915.)
- Freeman, Frank N. "Handwriting"; in *Sixteenth Yearbook of the National Society for the Study of Education*, part 1, pp. 60-72. (Bloomington, Illinois: Public School Publishing Company, 1917.)
- Freeman, Frank N. "Handwriting Scales for the Pupil"; in *Elementary School Journal*, vol. 21, pp. 755-61. (June, 1921.)
- Freeman, Frank N. "The Scientific Evidence on the Handwriting Movement"; in *Journal of Educational Psychology*, vol. 12, pp. 253-70. (May, 1921.)
- Gray, Clarence Truman. *A Score Card for the Measurement of Handwriting*. University of Texas Bulletin no. 37. (Austin: University of Texas, 1915.)
- Gray, Clarence Truman. "The Training of Judgment in the Use of the Ayres Scale for Handwriting"; in *Journal of Educational Psychology*, vol. 6, pp. 85-98. (February, 1915.)
- Johnson, Joseph Henry. "A Comparison of the Ayres and Thorndike Handwriting Scales"; in *The North Carolina High School Bulletin*, vol. 7, no. 4, pp. 170-73. (Chapel Hill: University of North Carolina, 1916.)
- Johnson, G. L., and Stone, C. R. "Measuring the Quality of Handwriting"; in *Elementary School Journal*, vol. 16, pp. 302-15. (February, 1916.)

- Kelley, T. L. "Comparable Measures"; in *Journal of Educational Psychology*, vol. 5, pp. 589-95. (December, 1914.)
- King, Irving, and Johnson, Harry. "The Writing Abilities of the Elementary and Grammar School Pupils of a City School System Measured by the Ayres Scale"; in *Journal of Educational Psychology*, vol. 3, pp. 514-20. (October, 1912.)
- Koos, L. V. "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools"; in *Elementary School Journal*, vol. 18, p. 422. (February, 1918.)
- Lewis, E. E. "The Present Standard of Handwriting in Iowa Normal Training High Schools"; in *School Administration and Supervision*, vol. 1, pp. 663-71. (December, 1915.)
- Lister, Clyde C. "The New York City Penmanship Scale"; in *Brooklyn Training School for Teachers Bulletin* no. 3. (Brooklyn, New York: Training School for Teachers, 1919.)
- Lister, C. C., and Myers, G. C. "An Analytic Scale of Handwriting"; in *Journal of Educational Psychology*, vol. 9, pp. 417-31. (October, 1918.)
- Manual, H. T. "The Use of an Objective Scale for Grading Handwriting"; in *Elementary School Journal*, vol. 15, p. 269. (January, 1915.)
- Mead, Cyrus D., and Welty, Howard O. "Practice in Using a Handwriting Scale"; in *University of California Studies in Elementary Education*, vol. 2, pp. 2-14. Bureau of Research in Education Study no. 9. (Berkeley: University of California, 1922.)
- Morton, R. L. "The Value of a Handwriting Scale to an Untrained Teacher"; in *Journal of Educational Research*, vol. 3, pp. 133-37. (February, 1921.)
- Nifenecker, E. A. "Grade Norms for the New York City Penmanship Scale"; in *Journal of Educational Research*, vol. 2, p. 809. (December, 1920.)
- Nutt, H. W. "Rhythm in Handwriting"; in *Elementary School Journal*, vol. 17, pp. 432-45. (February, 1917.)
- Pintner, R. "A Comparison of the Ayres and Thorndike Handwriting Scales"; in *Journal of Educational Psychology*, vol. 5, pp. 525-31. (November, 1914.)
- Reavis, W. C., and Aiken, N. J. "The Use of a Score Card in Measuring Handwriting"; in *Elementary School Journal*, vol. 19, pp. 36-40. (September, 1918.)
- Sackett, L. W. "Comparable Measures of Handwriting"; in *School and Society*, vol. 4, pp. 640-45. (October 21, 1916.)
- Starch, Daniel. "The Measurement of Efficiency in Writing"; in *Journal of Educational Psychology*, vol. 6, pp. 106-14. (February, 1915.)
- Starch, Daniel. "A Scale for Measuring Handwriting"; in *School and Society*, vol. 9, pp. 154-58, 184-88. (February 1, February 8, 1919.)
- Starch, Daniel. "A Revision of the Starch Writing Scale"; in *School and Society*, vol. 10, pp. 498-99. (October 25, 1919.)

204 EDUCATIONAL TESTS AND MEASUREMENTS

- Starch, Daniel. "Methods in Constructing Handwriting Scales"; in *School and Society*, vol. 10, pp. 328-29. (September 13, 1919.)
- Stone, C. R. "Motivation of the Formal Writing Lesson through a Special Classification of Pupils for Writing"; in *School and Home Education*. (June, 1915.)
- Thorndike, E. L., "Handwriting"; in *Teachers College Record*, vol. 2, no. 2. (March, 1910.)
- Thorndike, E. L. "Teachers' Estimates of Specimens of Handwriting"; in *Teachers College Record*, vol. 15, no. 5. (November, 1914.)
- Thorndike, E. L. "Handwriting Scales"; in *School and Society*, vol. 9, p. 236. (February 22, 1919.)
- Witham, Ernest. "The Most Accurate Measure of Handwriting"; in *Journal of Educational Administration and Supervision*, vol. 6, pp. 150-58. (June, 1920.)

CHAPTER V

SPELLING

I. THE PROBLEM OF MEASUREMENT IN SPELLING

Difficulties encountered. The measurement of spelling ability involves certain difficulties which are peculiar to the subject. Shall spelling ability be construed to mean ability to spell words when the attention is focused upon the ideas which are being expressed in writing, rather than upon the spelling of the words? Or is it the ability to spell words when the attention is focused upon the spelling of words, as in the case of dictated spelling lists? Upon the basis of what words shall one's spelling ability be measured? Shall spelling ability be defined in terms of the per cent of correct spellings of a limited group of frequently used words, or shall it be defined in terms of the extent of one's spelling vocabulary?¹

It appears to be the consensus of opinion that children should learn to spell correctly and with a minimum of attention the words used frequently in writing letters, compositions, school exercises, and the like. Otherwise, it is impossible for them to focus their attention upon the ideas which are being expressed. In addition, it is desirable that they should be able to spell a larger number of words which are used only occasionally. In the case of the more difficult and unusual of these words, it is probably sufficient if one is able to spell them correctly when attending to them.

¹ There are other phases of spelling ability, such as an ideal of correct spelling which will insure the use of the dictionary in the case of words concerning whose spelling a writer is uncertain, or the ability to detect words which are misspelled.

We shall consider first what the words most commonly used are, and how to measure the ability of pupils to spell them.

II. STANDARDIZED LISTS OF FOUNDATION WORDS¹

1. **Ayres Spelling Scale.**² In determining the most commonly used words, the method employed has been to examine written material of several types, such as letters, newspapers, and children's compositions, and to obtain a list of the words used and the number of times each word occurs. Ayres³ has combined the results of four such studies. Two of these studies were based on letters, the third upon newspapers, and the fourth upon selections of standard literature. The material examined in the four studies aggregated 368,000 words, written by twenty-five hundred different persons.

It was the original intention of Ayres to identify the two thousand most commonly used words, but he concluded that this was impossible because the material examined was found to consist of a few words used many times, and of a larger number of words used only a very few times. It was found that fifty different words were used so frequently that they made up approximately half of the material examined.

¹ Thorndike has recently published a list of 10,000 words. This is not intended to be used as a spelling list, but for certain purposes it may be valuable. See Thorndike, E. L., "Word Knowledge in the Elementary School"; in *Teachers College Record*, vol. 22, pp. 334-70. (September, 1921.) The list is published under the title, *Teacher's Word Book*, by the Bureau of Publications, Teachers College, Columbia University, New York City.

² This scale has been extended by B. R. Buckingham. The extension includes 505 new words, but they were not chosen in the same manner as Ayres chose his words, and hence should not be considered as belonging in a fundamental vocabulary in the same sense as the original scale. For the most part the added words are more difficult than those in the original scale.

³ Ayres, L. P., *Measurement of Ability in Spelling*. (Bulletin of the Division of Education, Russell Sage Foundation, New York City, 1915.)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
99	98	96	94	92	88	84	79	73	66	58	50	SECOND GRADE													
	THIRD GRADE	100	99	98	96	94	92	88	84	79	73	66	58	50	THIRD GRADE										
				FOURTH GRADE	100	99	98	96	94	92	88	84	79	73	66	58	50	FOURTH GRADE							
						FIFTH GRADE	100	99	98	96	94	92	88	84	79	73	66	58	50	FIFTH GRADE					
								SIXTH GRADE	100	99	98	96	94	92	88	84	79	73	66	58	50	SIXTH GRADE			
									SEVENTH GRADE	100	99	98	96	94	92	88	84	79	73	66	58	50	SEVENTH GRADE		
										EIGHTH GRADE	100	99	98	96	94	92	88	84	79	73	66	58	50	EIGHTH GRADE	
no	and	a	the	he	of	by	day	also	seven	became	catch	trust	except	eight	spend	sometimes	forenoon	often	guess	meant	principal	organization	immediate	decision	judgment
do	go	it	in	you	be	have	at	face	forget	brother	black	extra	am	am	enjoy	declare	noon	stopped	guess	earliest	testimony	emergency	convenient	principle	recommend
	at	is	so	will	but	are	sit	miss	happy	rain	warm	dress	capture	uncle	enjoy	engage	low	motion	earliest	whether	discussion	receipt	preliminary	disappoint	
	on	she	now	we	this	had	lot	ride	noon	keep	unless	beside	wrote	comfort	usual	complaint	combination	theater	argument	distinguish	arrangement	convenient	disappoint	allege	
		can	man	an	all	over	box	tree	think	start	clothing	happen	else	elect	auto	terrible	evening	improvement	volume	consideration	reference	emergency	disappoint		
		see	ten	my	your	must	door	sick	sister	mail	began	begin	offer	aboard	vacation	surprise	neighbor	century	organize	colonies	evidence	convenient	disappoint		
		run	bed	up	out	make	yes	got	card	glass	gone	file	suffer	shed	beautiful	period	vague	total	common	colonies	reference	convenient	disappoint		
			top	last	time	school	low	north	deep	upon	suit	provide	center	retire	flight	employment	various	mention	official	relief	convenient	disappoint			
				not	may	street	soft	white	inside	two	track	rule	front	refuse	travel	property	various	arrive	victim	occupy	convenient	disappoint			
				us	into	say	stand	spoke	blue	they	dash	stood	carry	district	trouble	select	various	supply	estimate	probably	convenient	disappoint			
				am	him	come	yard	black	post	any	fall	fix	royal	restrain	entrance	connection	various	assist	accident	foreign	convenient	disappoint			
				good	today	hand	bring	spring	town	could	buy	born	chain	object	importance	firm	various	examine	invitation	expense	convenient	disappoint			
				little	look	ring	fall	river	stay	only	stop	goes	death	pleasure	carried	region	various	particular	impossible	beginning	convenient	disappoint			
				ago	like	live	ball	plant	grand	about	walk	hold	learn	navy	loss	convict	various	affair	concern	application	convenient	disappoint			
				old	six	kill	law	cut	outside	city	grant	small	wonder	fourth	fortune	command	various	course	associate	various	convenient	disappoint			
				bad	boy	late	ask	water	game	where	soap	news	pair	population	proper	debate	various	marriage	associate	various	convenient	disappoint			
				red	book	let	just	stone	boat	week	war	summer	check	judge	mayor	public	various	serious	concern	various	convenient	disappoint			
						big	way	free	rest	first	above	enter	heard	weather	walt	represent	various	doubt	concern	various	convenient	disappoint			
						mother	get	lake	east	sent	express	turn	insight	worth	degree	term	various	condition	concern	various	convenient	disappoint			
						three	home	page	son	even	turn	lesson	itself	figure	engine	section	various	government	concern	various	convenient	disappoint			
						land	much	mile	help	without	turn	lesson	itself	figure	engine	section	various	opinion	concern	various	convenient	disappoint			
						old	call	end	hard	Friday	half	half	anything	forty	guest	relative	various	system	concern	various	convenient	disappoint			
						hat	long	fall	race	hour	father	state	high	instead	department	progress	various	piece	concern	various	convenient	disappoint			
						child	love	feet	cover	age	wife	gold	talk	personal	obtain	measure	various	witness	concern	various	convenient	disappoint			
						ice	then	west	fire	gold	read	fine	June	everything	family	serve	various	investigate	concern	various	convenient	disappoint			
						play	house	back	away	state	can	fine	right	rate	favor	estate	various	therefore	concern	various	convenient	disappoint			
						see	year	away	paper	July	head	fine	open	chief	husband	remember	various	purpose	concern	various	convenient	disappoint			
							to	put	each	story	head	fine	short	perfect	amount	effort	various	common	concern	various	convenient	disappoint			
							I	each	soon	open	head	fine	date	second	human	important	various	increase	concern	various	convenient	disappoint			
							as	some	came	line	head	fine	left	slide	view	duties	various	feature	concern	various	convenient	disappoint			
							send	some	show	show	head	fine	ship	farther	election	run	various	feature	concern	various	convenient	disappoint			
							one	how	Monday	yet	head	fine	train	duty	clerk	down	various	feature	concern	various	convenient	disappoint			
							has	her	yet	not	head	fine	water	intend	though	allow	various	feature	concern	various	convenient	disappoint			
							some	them	not	pay	head	fine	round	company	o'clock	position	various	feature	concern	various	convenient	disappoint			
							how	other	not	large	head	fine	cost	quies	support	held	various	feature	concern	various	convenient	disappoint			
							her	baby	not	near	head	fine	price	does	regard	ledge	various	feature	concern	various	convenient	disappoint			
							about	well	not	down	head	fine	power	wish	escape	claim	various	feature	concern	various	convenient	disappoint			
							men	for	not	why	head	fine	become	God	since	primary	various	feature	concern	various	convenient	disappoint			
							for	was	not	bill	head	fine	class	teacher	which	Saturday	various	purpose	concern	various	convenient	disappoint			
							his	that	not	girl	head	fine	care	November	length	appoint	various	therefore	concern	various	convenient	disappoint			
							led	lay	not	part	head	fine	try	April	destroy	information	various	purpose	concern	various	convenient	disappoint			
									not	still	head	fine	move	history	board	answer	various	purpose	concern	various	convenient	disappoint			
									not	place	head	fine	delay	study	September	reply	various	purpose	concern	various	convenient	disappoint			
									not	report	head	fine	ever	himself	station	oblige	various	purpose	concern	various	convenient	disappoint			
									not	never	head	fine	build	use	between	public									

All the words in each column are of approximately equal spelling difficulty. The steps in spelling difficulty from each column to the next are approximately equal steps. The numbers at the top indicate about what per cent of correct spellings may be expected among the children of the different grades. For example, if 20 words from column H are given as a spelling test it may be expected that the average score for an entire second grade spelling them will be about 79 per cent. For a third grade it should be about 92 per cent, for a fourth grade about 98 per cent, and for a fifth grade about 100 per cent.

The limits of the groups are as follows: 50 means from 46 through 54 per cent; 58 means from 55 through 62 per cent; 66 means from 63 through 69 per cent; 73 means from 70 through 76 per cent; 79 means from 77 through 81 per cent; 84 means from 82 through 86 per cent; 88 means from 87 through 90 per cent; 92 means from 91 through 93 per cent; 94 means 94 and 95 per cent; 96 means 96 and 97 per cent; while 98, 99 and 100 per cent are separate groups.

By means of these groupings a child's spelling ability may be located in terms of grades. Thus if a child were given a 20 word spelling test from the words of column O and spelled 15 words, or 75 per cent of them, correctly it would be proper to say that he showed fourth grade spelling ability. If he spelled correctly 17 words, or 85 per cent, he would show fifth grade ability, and so on.

Fig. 15. MEASURING SCALE FOR ABILITY IN SPELLING

Russell Sage Foundation, New York City
Division of Education
Leonard P. Ayres, Director

The data of this scale are computed from an aggregate of 1,400,000 spellings by 70,000 children in 84 cities throughout the country. The words are 1,000 in number and the list is the product of combining different studies with the object of identifying the 1,000 commonest words in English writing. Copies of this scale may be obtained for five cents apiece. Copies of the monograph describing the investigations which produced it may be obtained for 30 cents each, including the scale. Address the Russell Sage Foundation, Division of Education, 130 East 22d Street, New York City.

In order to secure a list of the thousand most frequently used words, it was necessary to include words which were found only forty-four times in the 368,000 words of material examined. This list of one thousand words is the best statement which we have of the words that form the core or foundation of the English language.

To determine the words of equal difficulty and the relative difficulty of the groups of words, Ayres divided the thousand words into fifty lists of twenty words each. Each list of words was spelled by the children of two consecutive grades in a number of cities. The thousand words were then divided into another fifty lists of twenty words each. Each of the new lists was spelled by the children in four consecutive grades. In all seventy thousand children spelled twenty words, making a total of 1,400,000 spellings, or an average of fourteen hundred spellings for each of the thousand words.

Upon the basis of this information Ayres classified the words into twenty-six groups, the words of each group being approximately equally difficult for school children of a given grade.¹ This classified list, together with the per cent of pupils in each of the grades who spelled the words of each list correctly, has been printed with the title, *Measuring Scale for Ability in Spelling*. This scale is reproduced as Fig. 15.² The per cents at the top of the columns of words are the standards or norms for a test made up of words taken from a column.

2. The Iowa Spelling Scales. In deriving the Iowa Spelling Scales, Ashbaugh followed much the same general plan as that just described. There are two significant differences. The written material examined in determining the

¹ For the details of the method employed see Ayres, L. P., *Measurement of Spelling Ability*, pp. 22-35. (Bulletin of the Division of Education, Russell Sage Foundation, New York City, 1915.)

² The author is indebted to Dr. Ayres for permission to reproduce this scale.

most commonly used words consisted of 3723 letters of Iowa people, totaling 361,184 words.¹ Ayres's list was based upon three types of material. The scales include a total of 2977 words instead of only one thousand. The scales are published in book form and hence are more expensive and less convenient to use than the Ayres Scale. The larger number of words, however, makes the list a more valuable one for many purposes.

3. **Second and third thousand most frequently used words.** A group of workers at Teachers College have prepared a list of the second and third thousand most frequently used words. In doing this they combined "list 5 of Eldridge" with "lists 1, 2, and 3 of Cook and O'Shea." "From the resulting list were eliminated of the Ayres's thousand, Jones's 'demons,' all the proper names of persons and places, all hyphenated words, all foreign words not found in a standard dictionary, and all compound words the parts of which were already listed in single forms unless changes in spelling or in pronunciation were found in the compounds." The two thousand words having the highest frequencies in the remaining list were taken as the second and third thousand most frequently used words. The words were classified according to difficulty on the basis of the spellings of high-school students.

III. SPELLING TESTS

The standardized lists of words which we have just described are not spelling tests. They are merely lists from which words may be selected for making a spelling test. Several workers have constructed spelling tests by taking words from these lists. Before these tests are described certain general matters relating to the making of spelling tests will be considered.

¹ This part of the work was done by W. N. Anderson.

Selecting words for a uniform test. It is a well-known fact that some words are more difficult to spell than others.¹ If the words for a test are taken at random from Ayres's list some will be easy and others relatively difficult. If equal credit were given for a correct spelling of the different words an error would be introduced in our measurements. Pupils spelling only the easier words correctly would receive a score higher than they deserved, and the bright pupils would receive a score lower than they deserved. This will tend to introduce errors in the measures of achievement in spelling. Therefore, if a uniform test is desired, the test words should be approximately equal in difficulty. Since the words in the three lists we have described have been classified according to difficulty, it is easy to select words which will be spelled correctly by approximately the same per cent of pupils.

How difficult words to use for a uniform test. In selecting the words for a spelling test, it is necessary to determine how difficult words should be used. When a pupil spells correctly all of the words of a given list, we do not have a measure of his spelling ability. We simply know that he can spell these words correctly; we do not have any information concerning how far beyond this list his spelling ability extends. In fact, the pupil has been given no opportunity to show how well he can spell. It is a well-known fact that the pupils of a typical grade or class are not equal in ability, but exhibit a wide range of ability. Thus, in testing a class in spelling it is necessary to use words for which the average per cent of correct spellings is less than one hundred. Otherwise we shall obtain a measure of the

¹ The spelling difficulty of a word has two interpretations. It may be taken to mean the difficulty which children experience in learning to spell it. It may also refer to the frequency with which it is misspelled. The latter meaning will be used in this chapter.

spelling ability of only the poorer pupils. Ayres recommends that in making a test for the pupils of a given grade, the words be taken from the column of his scale for which an average of eighty-four per cent of correct spellings may be expected.¹ Figure 16 represents a typical result of using

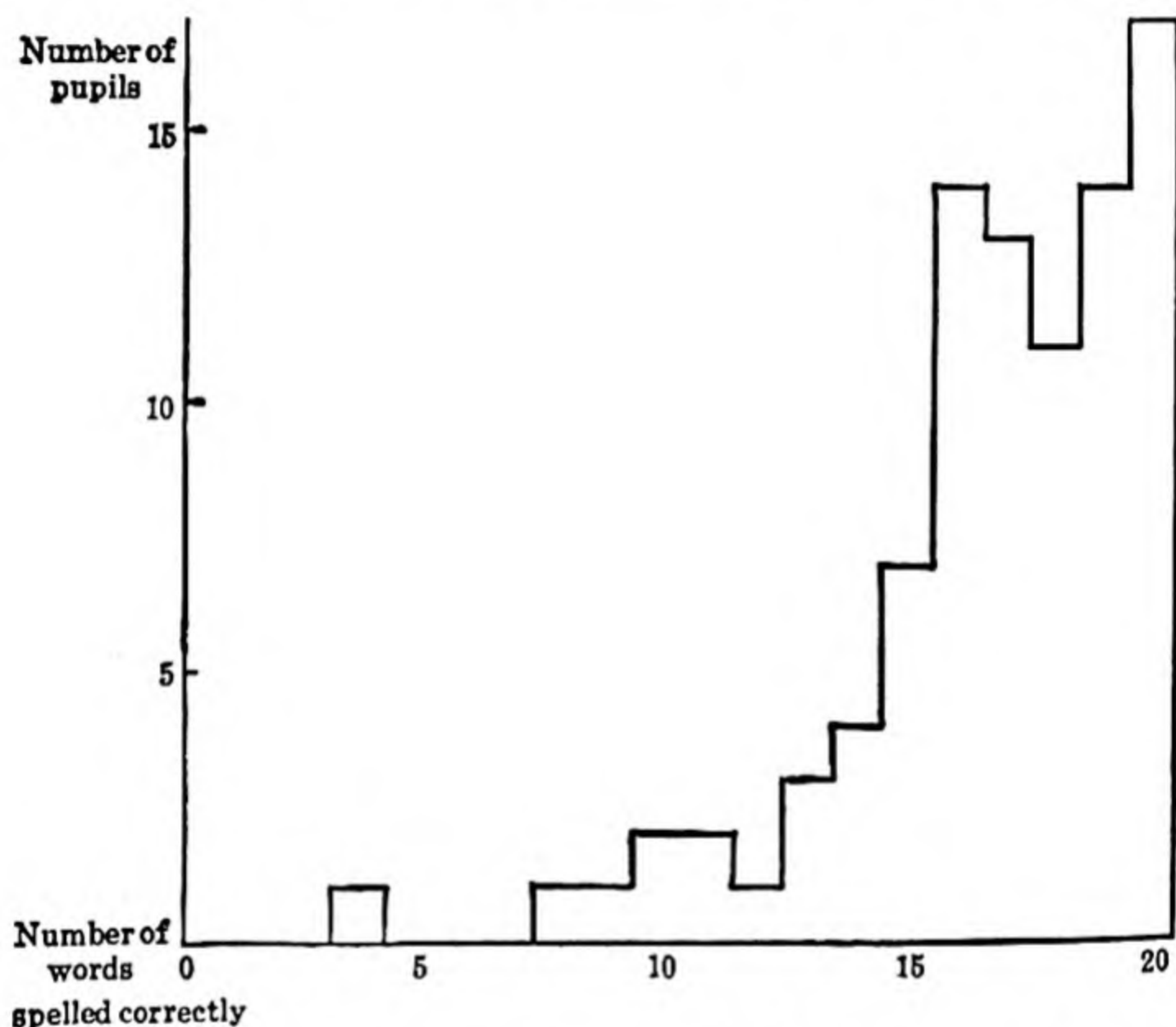


FIG. 16. SHOWING THE DISTRIBUTION OF 91 PUPILS ACCORDING TO THE NUMBER OF WORDS SPELLED CORRECTLY

Class average, 84 per cent

words chosen as Ayres recommends. The class average is eighty-four per cent, but those pupils who spelled all of the words correctly have not been tested. Those who mis-

¹ The reader should not confuse scores or measures of ability with school marks. The per cent of correct spellings is a measure. The school mark is the meaning which the school attaches to that measure. The fact that both the measure and the school mark may be expressed in per cents does not make them the same. See pages 7-9 for a more complete discussion of this point.

spelled only one or two words probably have not been tested satisfactorily.

Otis¹ presents facts from which he concludes that the most reliable measures of spelling ability are obtained by using words for which there is an average of fifty per cent of correct spellings. In support of this conclusion he points out that a list of words for which the average per cent of correct spellings was either zero per cent or one hundred per cent would yield a measure of zero reliability. Likewise a list of words for which the average per cent of correct spellings was ten per cent or ninety per cent would yield measures only slightly more reliable. Hence it seems natural that the most reliable measures would be obtained by using a list for which the average per cent of correct spellings was fifty. Thorndike has used words for which the per cent of correct spellings is fifty.² Ayres gives no satisfactory justification for recommending the choice of words for which an average of eighty-four per cent of correct spellings may be expected. When measuring the spelling ability of children in Springfield, Illinois, Ayres used words for which seventy per cent of correct spellings had been obtained. For the Survey of Cleveland, Ohio, the words were chosen from columns for which the average per cent of correct spellings was seventy-three.

In the sixth grade and above, the number of words in the columns of Ayres Scale for which the per cent of correct spellings is fifty, is so small that a satisfactory test cannot be made. Hence it is necessary to take words from an easier list or from two or more columns. The latter plan will not result in serious error, but for most purposes it will be satisfactory to use words from an easier list. If the Iowa Spell-

¹ Otis, A. S., "The Reliability of Spelling Scales"; in *School and Society*, vol. 4, p. 753.

² Thorndike, E. L., "Means of Measuring School Achievement in Spelling"; in *Educational Administration and Supervision*, vol. 1, p. 306.

ing Scales or the "second and third thousand most frequently used words" is used as a source for test words it will be possible to secure more difficult words. These lists should be used in constructing tests for the upper grades and high school unless one's purpose is to measure the ability to spell the words of the Ayres Scale.

In addition to being a list from which words may be taken for a spelling test, the Ayres Scale gives us the words which pupils should learn to spell with one hundred per cent accuracy. When this goal is attained, tests made from this scale will be of little value for measurement.

Selecting words for a scaled test. If it is desired to construct a test in the form of a scale, one may take one word from each column of the Ayres Scale. More difficult scaled tests may be formed by selecting words in a similar manner from the other two lists. A measuring instrument of this type has certain advantages. It can be used in all grades instead of only one or two grades as is the case of uniform tests. A scaled spelling test is useful for survey purposes. If all of the words are taken from Ayres's list, the criticism of this type of measuring instrument mentioned on page 55 has only a limited application. The number of words spelled correctly is usually taken as the pupil's score on this type of test.¹

How many words to use. Another question which must be considered in making a spelling test is the number of words it is necessary to use. In general the ability to spell one word is separate and distinct from the ability to spell any other word. Ability to spell, therefore, consists of a large number of abilities to spell specific words. This being the case it would be necessary to use all of the thousand words of Ayres's list in order to obtain a complete and accurate measure of a pupil's ability to spell the most com-

¹ For a comparison of weighted and unweighted scores see Monroe, Walter S., *An Introduction to the Theory of Educational Measurements*, p. 128. (Houghton Mifflin Company, Boston, 1923.)

monly used words. However, it is possible to secure a measure which is representative of the pupil's ability to spell these words by using a smaller number of words. This is possible in just the same way that it is possible to determine the quality of a load of wheat or a vat of cream by the examination of a sample.

How many words are necessary in making a spelling test depends upon what is desired. Relying upon the theory of random sampling, Thorndike believes a small number of words is sufficient to measure the spelling achievement of a large school system. A test consisting of only ten words has been used in a number of school surveys. This number is probably sufficient for the measure of a large school system, but if it is desired to obtain a measure of the spelling ability of individual pupils, a larger number must be used. Otis¹ says that a twenty-five word test gives a very poor measure of individual ability, and that at least one hundred words should be used, better four hundred or five hundred words. Starch recommends the use of two hundred words. In view of the time required for marking the test papers, it is probably wise to limit a test to fifty words.

Methods of giving the test. The words which make up a test may be dictated to the pupils as separate words, or they may be embedded in sentences which are dictated. Furthermore, the dictation of the sentences may be timed so that the pupils will approximate their normal rate of writing. Investigation has shown that the per cent of correct spellings is higher when the words are dictated separately than when they are dictated in timed sentences. According to Courtis the per cent of correct spellings is about five greater when the words are dictated in lists. Fordyce has found this difference to be between ten and fifteen per cent.

This means that the ability to spell words when the attention is focused upon the spelling, as is the case when the

¹ *Loc. cit.*, pp. 679, 682.

words are dictated separately, is not the same as the ability to spell the same words when the act of spelling is in the margin of one's attention. In writing letters, compositions, and the like, the spelling must be carried on in the margin of the attention because the ideas which are being expressed must occupy the focus of the attention. This is particularly true of the foundation words of the language such as we have in the Ayres list. The words of this list constitute over ninety per cent of the words we use. Hence, by using the words embedded in sentences which are dictated so that the time allowed corresponds to the average rate of writing, we are more likely to secure a measure of that spelling ability which functions in the writing of letters and school exercises.

Letters per minute. Pupils may be caused to write at approximately their normal rate by dictating the sentences at that rate. One set of norms for rate of handwriting are as follows in terms of letters per minute: second grade, 36 letters; third grade, 48 letters; fourth grade, 56 letters; fifth grade, 65 letters; sixth grade, 72 letters; seventh grade, 80 letters; eighth grade, 90 letters. The dictation of a sentence requires some additional time, probably 10 per cent. For example, in the case of the sixth grade, instead of dictating at the rate of 72 letters in one minute, 66 seconds should be allowed for words totaling 72 letters. On the basis of the above norms the number of seconds to be allowed per letter for the several grades are as follows:

GRADE	SECONDS PER LETTER
II.....	1.83
III.....	1.38
IV.....	1.18
V.....	1.01
VI.....	.92
VII.....	.83
VIII.....	.73

If the sentences contain more than thirty to forty letters, they should be dictated in sections, so that the pupils' writing will not be slowed up by trying to recall what has been dictated. Furthermore, tests of rate of handwriting have shown that all pupils do not normally write at the same rate. For this reason provision must be made for those pupils who are accustomed to write more slowly than the standard rate. This can be done by having none of the test words come at the end of the sentences, and requiring all pupils to begin upon the next sentence as soon as it is dictated, even if they have not finished writing the preceding.

Summary. The teacher should use the lists described on pages 206-208, as sources of words for constructing spelling tests. These tests should be constructed according to the following principles, which have been considered in the preceding pages:

1. The words for a uniform test should be chosen so that they will be approximately equal in spelling difficulty.
2. A scaled test may be constructed by taking one word from each column.
3. Twenty words are probably sufficient to secure a reliable measure of the spelling ability of a class. At least fifty words should be used to secure a reliable measure of the spelling ability of individual pupils. In the case of the upper grades it will be necessary to use words from more than one column when using the Ayres Scale.
4. In order that the words may be difficult enough to measure the spelling ability of all pupils, they should be chosen from columns for which the standard per cent of correct spellings is not more than seventy. For the lower grades it is probably best to use words for which the standard per cent of correct spellings is from fifty to sixty-six. If the words are to be used in timed sentences it will probably be satisfactory to use slightly easier words.
5. The words should be embedded in sentences, and the sentences dictated at approximately the standard rate of hand-

writing for the grade. Test words should not occur at the end of the sentences.

Directions for giving a timed sentence test. The following test has been constructed in accordance with the above principles. The directions given below should be followed in giving it:

1. See that the pupils are provided with two or three sheets of paper, and with either pencil or pen and ink. If pencils are to be used, they should be well sharpened. If pen and ink are used, good pen points should be provided.
2. Say to the pupils: "I have some sentences which I want you to write as I dictate them. I am going to dictate them rather rapidly, possibly more rapidly than some of you can write. If you have not finished writing one sentence when I begin to dictate another, I want you to leave it and begin on the new sentence. If there are any words you cannot spell you may omit them. Take time to dot your *i*'s and cross your *t*'s. If you have any question about what you are to do, ask it now because you cannot ask questions after I begin to dictate."
3. Make certain that all pupils understand what they are to do. It is well to give a short preliminary practice in writing from dictation if the pupils are not accustomed to it. For this purpose use some simple selection, but dictate it at the standard rate.
4. Dictate the first sentence when the second hand of your watch is at 60. When it reaches 27,¹ dictate the second sentence. When it reaches 13, dictate the third, and so on. Dictate the sentences distinctly, but do not repeat. It is advisable for the teacher to practice dictating the sentences according to the directions before attempting it with a class.
5. Stop the pupils promptly at the time indicated. Allow no corrections to be made. Ask the pupils to turn their papers over and write their name and grade. Collect the papers.

¹ This position of the second hand is for the timed sentence spelling test illustrated below. The position would probably be different for another test. The same is true of the third position, "13."

A TIMED SENTENCE SPELLING TEST OF FIFTY WORDS TAKEN FROM
COLUMN O

(Arranged for a Fourth Grade)

- (60) The *public* appear to want it.
- (27) The *population* of the *district* is five hundred.
- (13) We *refuse* to attend the meeting.
- (44) My *uncle* will remain until the man comes.
- (23) *Forty* members of the *fourth* company will drill.
- (9) The *judge* knew the *chief* of *police* was there.
- (52) The *whole* address was tiresome.
- (23) The *comfort* of our *friend* is to be considered.
- (7) A *perfect* figure was drawn.
- (33) During the month of *August* we *elect* a teacher.
- (17) The *navy* will go farther away.
- (45) A *board* from the *shed* is needed.
- (15) The house was *between* the *jail* and the store.
- (58) I am *getting* the *second* order.
- (26) The *station* is worth more money.
- (57) *Madam* will return *Thursday* morning.
- (32) We *don't* request attendance.
- (60) What is the *objection* to a *personal* letter?
- (41) A *sudden* change in the *weather* will come soon.
- (25) *Duty* before *pleasure* is an old saying.
- (2) *Eight* men *intend* to retire from the house.
- (42) It would be *proper* to call.

When the second hand reaches 7, stop the work.

Grade norms. In classifying the words of his list according to difficulty, Ayres determined the average per cent of the pupils of each grade who spelled the words correctly. Thus the words of column O (see Fig. 15) were spelled correctly by 50 per cent of the third-grade pupils, 73 per cent of the fourth-grade pupils, 84 per cent of the fifth-grade pupils, 92 per cent of the sixth-grade pupils, 96 per cent of the seventh-grade pupils, and 99 per cent of the eighth-grade pupils. These per cents, which are printed at the head of each column, represent the average spelling ability of pupils in the

several grades when the words are dictated in lists. When the words are used in timed sentences the per cent of correct spellings has been 5 to 15 per cent lower.

In Boston minimum word lists for each grade have been carefully built up on the basis of the words which the pupils use in their written work. When the Boston pupils were tested on the words of the Ayres Scale which are included in their several grade lists, the per cent of correct spellings was conspicuously above the grade norms given by Ayres.¹ In reporting that study Ballou suggests that "this may be due to the fact that the Boston pupils had been taught the words; whereas, the pupils in the eighty-four cities where Ayres gave his lists, and on the results of which he standardized the words for his spelling scale, may not have been taught them."

For this reason it may be seriously questioned whether the averages which Ayres gives are satisfactory grade norms of spelling ability for the foundation words of the language. Ayres says: "Probably the scale will have served its greatest usefulness in any locality when the school children have mastered these one thousand words so thoroughly that the scale has become quite useless as a measuring instrument." In the past we have not had the advantage of such a list and have distributed our efforts in teaching spelling over a very much larger list of words. If we accept these one thousand words as the foundation words of our language, we should place prime emphasis upon teaching them. This being the case a satisfactory eighth-grade norm would approximate one hundred per cent for all of the words. For the preceding grades, the norms would be one hundred per cent for the words of the list which the pupils had been taught. For example, the easiest nine hundred words might be used for

¹ Ballou, F. W., "Measuring Boston's Spelling Ability by the Ayres Spelling Scale"; in *School and Society*, vol. v, pp. 267-70.

the seventh grade, the easiest seven hundred and fifty for the sixth grade, and so forth. The use of the scale in the way Ayres suggests would seem to lead to grade norms of this type. The distribution of the words among the several grades and the optimum grade norms must be determined by experimentation.

Ashbaugh gives similar grade norms for the Iowa Spelling Scales, but since a much larger number of words is included, the suggested modification of the norms for the Ayres Scale would apply with less completeness. In the case of the words which occur least frequently, it is possible that norms of less than one hundred per cent would be satisfactory. In the case of the Teachers College list, grade norms are not given.

1. Monroe Timed Sentence Spelling Tests

General structure. In order to make easily available a timed sentence spelling test, the writer constructed a series of such tests, using test words chosen from appropriate columns of Ayres's Scale for Measuring the Ability in Spelling and basing the rate of dictation upon the measurements of the rate of handwriting of over six thousand Kansas school children. In order that the scores might have a high degree of reliability as measures of the spelling ability of individual pupils, fifty test words were used in each test. According to one study¹ the probable error of an individual score for a test of fifty words is less than 1.00 when the score is expressed as the per cent of words spelled correctly. For a class of twenty-five or more pupils the probable error of the class score would be 0.2.

For Grades III and IV the test words were taken from

¹ Otis, A. S., "The Reliability of Spelling Scales Involving a 'Deviation Formula' for Correlation"; in *School and Society*, vol. 4, pp. 716-22. (November 11, 1916.) This study deals with the reliability of tests consisting of isolated words and it is possible that the results might not apply to a timed sentence spelling test.

Column M of Ayres's Scale for Measuring Ability in Spelling. For Grades V and VI they were taken from Column Q and for Grades VII and VIII and the high school they were taken from Columns S, T, and U. In these tests no test words come at the end of the sentences. Thus, the pupil who writes slowly will be much less likely to make low scores because he does not have time to complete writing the sentences. It should also be noted that all other words found in these sentences are easier to spell as shown by the Ayres Scale.

Norms. This series of timed sentence spelling tests was given in sixteen Kansas cities in April and May, 1917. During the school year of 1919-20 scores were reported from a number of cities. The grade medians for these two groups of cities and the norms given by Ayres are given in Table XXII. In comparing the successive grades it must be remembered that the same test words were not used for all grades. One list of test words was used for Grades III and IV, another for Grades V and VI and still another for Grades VII and VIII and for the high school.

The fact that most of the median scores in Table XXII are materially below Ayres's norms indicates that a different type of spelling ability has been measured by the timed spelling sentence test than that measured by Ayres in constructing his scale. (Ayres had the words dictated in lists.) This fact becomes more apparent when it is recalled that many of the cities which gave these tests had used Ayres's Scale as a minimum course of study as well as a source of test words. Thus, had the test words been dictated in lists it is likely that the median scores would have been materially above Ayres's norms.¹

¹ It is possible that the difference between the median scores and Ayres's norms may be due to factors other than the measurement of different types of spelling ability. Many of the pupils probably were not accustomed to

TABLE XXII. MEDIAN NUMBER OF WORDS SPELLED CORRECTLY FOR MONROE TIMED SENTENCE SPELLING TESTS — TOTAL WORDS IN EACH TEST IS FIFTY

	GRADE									
	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1917 TESTING										
Number of pupils	1342	1539	1427	1338	1136	876	289	251	188	96
Median (May)	28.6	39.4	32.8	41.8	35.1	43.0	43.3	45.1	46.5	47.9
1919-20 TESTING										
Number of pupils	437	423	440	302	218	153				
Median	28.2	43.0	39.7	43.0	32.6	38.8				
Ayres's Norms	33.0	42.0	36.5	42.0	39.0	44.0				

2. Courtis Standard Dictation Spelling Tests

These tests are very similar to the Monroe Timed Sentence Spelling Tests just described. There are two forms for each Grade from III B to VIII A. Each test includes twenty-five test words.

3. Seven S Spelling Scales

General structure. The Seven S Spelling Scales are a group of "Sixteen Spelling Scales Standardized in Sentences for Secondary Schools." The words for these scales were selected from the second and third thousand most frequently used words. (See page 208.) These tests are in scale form and each consists of twenty short sentences. Each sentence includes only one test word, which is italicized. The sentences are not timed. The directions for using the scales state: "Read the sentence aloud to the pupils; then pronounce the italicized word to denote the word they are to

writing from dictation and all were not accustomed to writing at the rate at which these tests were dictated. It is possible that these unusual conditions may have been operated to materially lower the scores of a number of pupils or even of most pupils.

spell. Read the sentence and repeat the word again." The first twelve of the scales are to be used as duplicate forms. The last four form another group of equivalent scales but are more difficult than the first twelve.

Norms. Norms in terms of per cent of correct spellings and based upon February testing are given below. To calculate the norms for any other month the following procedure is recommended. "Counting ten months to the school year, standards for any school month may be derived by allowing for each month one tenth of the difference between the standards for the two years between which that month falls."

GRADE	SCALES I-XII	SCALES XIII-XVI
VII.....	65.90	34.76
VIII.....	73.77	45.03
IX.....	80.00	53.91
X.....	85.05	61.48
XI.....	88.67	67.08
XII.....	91.25	72.14

4. Starch's Spelling Scale

Measuring the extent of a pupil's vocabulary. In securing a measure of the number of words which a pupil or a class can spell correctly, we are not concerned simply with the most commonly used words of the English language, but rather with all the words of the language.¹ Starch has prepared a set of six word lists, each consisting of one hundred words. The words for these tests were chosen by taking the first defined word on the even-numbered pages in Webster's New International Dictionary (1910 edition). Technical, scientific, and obsolete words were discarded

¹ Starch, Daniel, "The Measurement of Efficiency in Spelling, and the Overlapping of Grades in Combined Measurements of Reading, Writing, and Spelling"; in *Journal of Educational Psychology*, vol. 6. (March, 1915.)

from the list. The remaining six hundred words were arranged alphabetically according to size. When arranged in this way they were divided into six lists of one hundred words each by taking the first, seventh, thirteenth, and so forth for the first list; the second, eighth, fourteenth, for the second list; and so forth. Each of these lists consists, therefore, of one hundred words taken at random from the non-technical words of the English language. Such a list is an instrument for measuring the extent of a pupil's correct spelling vocabulary.

The words in each list are arranged according to the number of letters they contain. Ayres¹ found for the words of his list a very high correlation between the length of words and spelling difficulty. Assuming this to be true for Starch's list, the words may be considered to be arranged approximately in the general order of spelling difficulty by groups. A pupil's score is the number of words spelled correctly, no account being taken of the difficulty of the word spelled. The score is an index of the total number of the non-technical words of the English language which the pupil can spell correctly. It may be pointed out that in general this type of test is less useful than those which we have already described. Its norms should not be emphasized as educational objectives. The major emphasis should be placed upon the words which are most frequently used, but some attention should be given to the size of pupils' spelling vocabulary.

Starch spelling lists. Since the spelling tests devised by Starch are not easily obtained lists I and II are given on the following pages. In using these lists as tests the words are simply dictated, and the pupil is thus allowed to focus his attention upon the spelling. To secure a reliable measure

¹ Ayres, L. P., *Measurement of Spelling Ability*, p. 38. (Bulletin of Division of Education, Russell Sage Foundation, New York City, 1915.)

224 EDUCATIONAL TESTS AND MEASUREMENTS

Starch recommends that both lists be used and the two sets of scores averaged.

STARCH'S SPELLING LIST, No. 1

- | | | |
|-----------|--------------|---------------------|
| 1. add | 35. prism | 69. commence |
| 2. but | 36. rogue | 70. estimate |
| 3. get | 37. shape | 71. flourish |
| 4. low | 38. steal | 72. luckless |
| 5. rat | 39. swain | 73. national |
| 6. sun | 40. title | 74. pinnacle |
| 7. alum | 41. wheat | 75. reducent |
| 8. blow | 42. accrue | 76. standing |
| 9. cart | 43. bottom | 77. venturer |
| 10. come | 44. chapel | 78. ascension |
| 11. easy | 45. dragon | 79. dishallow |
| 12. fell | 46. filter | 80. imposture |
| 13. foul | 47. hearse | 81. invective |
| 14. gold | 48. laden | 82. rebellion |
| 15. head | 49. milden | 83. scrimping |
| 16. kiss | 50. pilfer | 84. unalloyed |
| 17. long | 51. rabbit | 85. volunteer |
| 18. mock | 52. school | 86. cardinally |
| 19. neck | 53. shroud | 87. connective |
| 20. rest | 54. starch | 88. effrontery |
| 21. spur | 55. vanity | 89. indistinct |
| 22. then | 56. bizarre | 90. nunciature |
| 23. vile | 57. compose | 91. sphericity |
| 24. afoot | 58. dismiss | 92. attenuation |
| 25. black | 59. faction | 93. fulminating |
| 26. brush | 60. hemlock | 94. lamentation |
| 27. close | 61. leopard | 95. secretarial |
| 28. dodge | 62. omnibus | 96. apparitional |
| 29. faint | 63. procure | 97. intermissive |
| 30. force | 64. rinsing | 98. subjectively |
| 31. grape | 65. splashy | 99. inspirational |
| 32. honor | 66. torpedo | 100. ineffectuality |
| 33. mince | 67. worship | |
| 34. paint | 68. bescreen | |

STARCH'S SPELLING LIST, No. 2

- | | | |
|-----------|--------------|---------------------|
| 1. air | 35. quill | 69. covenant |
| 2. cat | 36. rough | 70. eugenics |
| 3. hop | 37. shout | 71. friskful |
| 4. man | 38. stick | 72. luminous |
| 5. row | 39. swear | 73. opulence |
| 6. tap | 40. trump | 74. planchet |
| 7. awry | 41. whirl | 75. reformer |
| 8. blue | 42. action | 76. thorough |
| 9. cast | 43. bridle | 77. watering |
| 10. corn | 44. charge | 78. belonging |
| 11. envy | 45. driver | 79. displayed |
| 12. feud | 46. finger | 80. indention |
| 13. game | 47. heaven | 81. mercenary |
| 14. grow | 48. legend | 82. redevelop |
| 15. home | 49. motley | 83. senescent |
| 16. knee | 50. portal | 84. uncharged |
| 17. look | 51. recipe | 85. whichever |
| 18. mold | 52. scrape | 86. centennial |
| 19. part | 53. simple | 87. constitute |
| 20. ruin | 54. strain | 88. exaltation |
| 21. take | 55. weaken | 89. invocative |
| 22. tree | 56. breaker | 90. personable |
| 23. well | 57. congeal | 91. strawberry |
| 24. allay | 58. disturb | 92. concentrate |
| 25. blaze | 59. foreign | 93. imaginative |
| 26. buggy | 60. hoggerly | 94. mathematics |
| 27. clown | 61. meaning | 95. selfishness |
| 28. doubt | 62. onerate | 96. collectivity |
| 29. false | 63. provoke | 97. marriageable |
| 30. forth | 64. salient | 98. agriculturist |
| 31. grass | 65. station | 99. quarantinable |
| 32. house | 66. trample | 100. relinquishment |
| 33. money | 67. abstract | |
| 34. paper | 68. bulletin | |

Norms. Starch gives the following standards for his tests based on their use with over twenty-five hundred pupils.

	GRADE							
	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>
Per cent of words spelled correctly	10	30	40	51	61	71	78	85

These standards are interpreted thus: the average eighth-grade pupil should be able to spell correctly eighty-five per cent of the non-technical words of the English language, or eighty-five of the one hundred words in any one of Starch's tests.

IV. INTERPRETATION OF SCORES AND REMEDIAL INSTRUCTION

General interpretation. In interpreting the general condition of a class one may follow a procedure similar to that described for arithmetic on page 76. Since only one score is secured for each pupil certain cases stated for a test which yields two scores will not arise. In interpreting a low class average in terms of instructional needs it is necessary to remember that it may be due to one or more of three conditions:

1. The class as a whole may be unable to spell certain words.

2. Certain pupils may be unable to spell a large number of the words of the test.

3. The errors may be rather uniformly distributed as to both words and pupils.

To determine the extent to which each condition causes

the low class average, the teacher should make the following type of tabulation from the test papers.

WORDS OF THE TEST	PUPILS											
	1	2	3	4	5	6	7	8	9	10	11	12
catch.....	c	-	-	-	c	-	-	-	-	c	-	c
black.....	c	c	c	c	-	c	c	c	c	c	-	c
warm.....	c	c	c	c	c	c	c	c	c	c	c	c
unless.....	c	c	c	c	c	c	c	c	c	c	-	c
clothing.....	c	-	c	-	-	c	-	-	c	c	-	-
began.....	c	c	c	c	c	c	c	c	c	c	-	c

c indicates the word was correctly spelled.

Although these words are listed by Ayres as being equally difficult for pupils in general, they are not necessarily so for particular pupils. Obviously in the class here represented "catch" and "clothing" need general emphasis, while only certain pupils need to give attention to "black," "began," and "unless." Pupil 11 has misspelled five out of six words, and hence probably is a "poor speller."

Individuality in spelling difficulties. Simply to know that a pupil is below standard in ability is of little value to the teacher, because spelling ability is specific and not general. In general the ability to spell one word does not imply ability to spell another word, nor does the lack of ability to spell a given word indicate that a pupil cannot spell another word. Hence the teacher should make a very careful diagnosis of the spelling ability of each pupil whose test score is below standard to ascertain just what words he cannot spell of those he is expected to spell.

This is accomplished by giving the pupils below standard a test including all of the words which they are expected to

be able to spell. Such a test is not for the purpose of measurement but should be thought of as the first step in the teaching of spelling. Each pupil should be required to make from this test a list of all the words which he has spelled incorrectly. The words of this list are the ones he needs to study. It is obvious that to ask a pupil to study words which he can already spell correctly is to ask him to use his time without profit.

"Spelling Demons." Certain frequently used words are very frequently misspelled. Jones¹ has given us a list of 100 words which he found misspelled most frequently in children's compositions. He calls them the "One Hundred Spelling Demons of the English Language." Nine tenths of these words are found in Jones's list for the second and third grade. Four fifths of these words are found in Ayres's list. A teacher will make no mistake in emphasizing these words in the teaching of spelling until the pupils can spell them correctly.

THE ONE HUNDRED SPELLING DEMONS OF THE ENGLISH LANGUAGE
(Jones)

which	can't	guess	they
their	sure	says	half
there	loose	having	break
separate	lose	just	buy
don't	Wednesday	doctor	again
meant	country	whether	very
business	February	believe	none
many	know	knew	week
friend	could	laid	often
some	seems	tear	whole
been	Tuesday	choose	won't
since	wear	tired	cough
used	answer	grammar	piece
always	two	minute	raise

¹ Jones, N. Franklin, *Concrete Investigations of the Materials of English Spelling*. (University of South Dakota Bulletin, 1913.)

where	too	any	ache
women	ready	much	read
done	forty	beginning	said
hear	hour	blue	hoarse
here	trouble	though	shoes
write	among	coming	to-night
writing	busy	early	wrote
heard	built	instead	enough
does	color	easy	truly
once	making	through	sugar
would	dear	every	straight

Types of misspellings. A pupil's spelling difficulty is not completely diagnosed when the words he does not spell correctly are located. Errors in spelling are seldom if ever distributed uniformly among the several letters composing the word. Neither does it appear that there is much uniformity in the location of errors in different words. Certain words are misspelled in only a few ways, while other words are misspelled in many ways. Certain misspellings occur frequently, while others seldom occur. In Table XXIII the misspelling of certain words found in the papers of 80 seventh-grade pupils are given, together with the frequency of each. The words were taken from column S of the Ayres Scale. Where no number follows the word that type of misspelling of the word occurred but once.¹

TABLE XXIII. THE MISPELLINGS OF EIGHTY SEVENTH-GRADE PUPILS IN A COLUMN SPELLING TEST

I. affair	II. assist	
affere	assit (3)	asscest
affire	asist (2)	assiaist
afair (2)	ascist	acsist
affaired	asest	acist (2)
affer	assaist	accisted
		assantant

¹ See Monroe, Walter S., *Report of Division of Educational Tests for 1919-20*. (Bureau of Educational Research Bulletin No. 5; University of Illinois Bulletin, vol. 18, no. 21.) Chapter IV of this Bulletin reports a more elaborate investigation of types of misspellings for certain words.

	assised	marrage (5)	perticular (8)
	accessise	marage	particular
	accest	merriage	pertucular
	astist	X. mention	particular
	assis	mension (8)	partiular
	assite	mentioned	partular
III. certain		meantion (2)	paticular
	certian (7)	menchion	perciluar
	serten	XI. motion	pectuliar
	sertain	moshen	pecticular
	certain	moticem	pertictural
	secrtain	motation	patuclure
IV. difference		montion	peculiar
	differance (10)	XII. neither	pectulair
	diffierence	neather (6)	XV. possible
V. examination		nether	possable (4)
	examation (10)	niether (2)	posible
	examition	nieghter	posable
	examnition	XIII. opinion	posiable (2)
	excamation	oppinion (5)	possiable (5)
	excanitions	opinon (2)	posobile
	examanation (3)	opinton	possibbe
VI. government		oppoinen	posiple
	goverment (9)	oppinum	XVI. serious
	govament	oppenion (2)	cyreaus
	gouvernement	opion (3)	cerrious
	gorvement	oponion (2)	scerious
VII. improvement		oppion (2)	serrious (2)
	improvment (7)	opininion	cerious
	impovement	opoin (2)	sereaus
VIII. investigate		opionion	XVII. stopped
	investagate (3)	XIV. particular	stoped (13)
	envesigatige	particular	stopst
	investiage	particuler	stocted
IX. marriage		partictuler	stop

Teaching the pupil to correct his errors in spelling. Spelling consists in forming correct and fixed associations "between the successive letters of a word and between the word thus spelled and the meaning."¹ The laws governing the formation of fixed associations are those of habit formation.

¹ Freeman, F. N., *Psychology of the Common Branches* (Houghton Mifflin Company), p. 115.

The first step in habit formation is to get the attention of the child focused upon the associations to be formed. The second step is to secure sufficient repetition. Repetition of the associations is secured both through drill and through using the word in written expression. The pupil must give attention to the repetitions of the associations in order to insure that wrong associations will not be made.

Numerous experiments have shown that pupils can spell correctly a large per cent of the words in the lists in spellers before they have studied them. Because of this fact the assignment of the spelling lesson should include the dictation of the words to the pupils so that each may know what words he needed to study. By this procedure the teacher will learn what words she should emphasize in her instruction.

Some writers state that a pupil should not be permitted to spell a word incorrectly when it can be avoided, and for this reason pupils should learn to spell words correctly before they are required to write them. Just how important it is to do this we do not know. In certain cases it appears that a child or an adult learns to spell certain words correctly by having his attention directed to his errors. The fact of his error serves to direct his attention to learning to spell the word correctly. Those who believe that evil effects will come from having pupils write words which they cannot spell correctly may direct them to omit those words which they think they cannot spell correctly.

The dictation of the words in assigning the spelling lesson, together with the detailed testing of the pupils as suggested on page 227, reveals to the teacher the words upon which she must exercise her ability as a teacher of spelling. It also reveals to her the pupils to whom instruction should be directed in the case of each word. Particular methods and devices by means of which the laws of habit formation may

be fulfilled are described in books which deal with the teaching of spelling.¹

Causes of some misspellings. Certain associations in the spelling of a word appear to be more crucial than others. Take, for example, the word "examination," the fifth word in Table XXIII. The letters *e-x* and *t-i-o-n* were correctly associated in every instance. All of the errors occurred in the other syllables. A study of Table XXIII shows that certain forms of misspelling occur more frequently than others, and that most of the misspellings may be attributed to certain specific causes. Forms of misspelling such as "partiular," "partuler," "opinon," "impovement," "pos-sibbe," are probably due to carelessness or accident. Relatively few of the misspellings in Table XXIII may be assigned to this cause. Errors of this type probably cannot be entirely eliminated from uncorrected manuscript. However, drill will reduce the number of such errors to a satisfactory minimum.

A more prolific source of error is mispronunciation of the word by the pupil. He may have acquired this from the teacher, but more likely from those with whom he associates outside of school. Or it may have been acquired from lack of attention to the form of the word. Such misspellings as the following are probably caused by mispronunciation: "peticular," "particlar," "investagate," "goverment," "examation."

A very striking instance of this type of spelling error and its cause came to the attention of the writer a few years ago. A man who had taught geometry for a number of years used the word "frustum" in a manuscript, spelling it "frustrum" which agreed with his pronunciation of the word. This manuscript was read by a number of well-known mathe-

¹ A very good chapter (vi) will be found in Freeman's *Psychology of the Common Branches*. See also Cook and O'Shea, *The Child and his Spelling*.

maticians who read it critically. Only two noted the misspelling of the word, and one mathematician, who took much pride in his ability to spell correctly and who was the author of several textbooks, admitted that he had always pronounced and spelled the word "frustrum."

Other errors listed in Table XXIII are due to certain phonic irregularities of the English language, for example, certain misspellings of "assist," "certain," "affair," "marriage," "motion," "neither," and "serious." Still other errors, such as, "stoped," and "improvment," are due to certain silent letters. In a few cases it appears that the pupil was not acquainted with the form of the word. Such cases probably should not be counted as errors but rather as words unknown.

Good teaching of spelling. In teaching the spelling of a word the child's attention should be directed to the crucial associations. If the word is one like "government," his attention should be called to the correct pronunciation. If it is such a word as "their," his attention should be called to the use of the word. To eliminate spelling errors a pupil's attention should be called to his particular error and he should be helped to remove the cause. If the cause is mispronunciation, see that he learns to pronounce the word correctly. If the error is due to a confusion of letters the pupil should be given some device to prevent this confusion. The following is a device which may be used for especially difficult words:

Par-tic-u-lar

I frequently misspell ——— in writing compositions but now I am going to learn to spell it correctly. My teacher tells me that I do not look at the ——— syllables and letters closely enough. I am going to do it now with ——— care. I see that the word has ——— syllables. The first syllable is ———. The vowel of this syllable is ———, the first letter of the alphabet. The last syllable

ble is ——— and the vowel is also ———. The word contains ——— letters, the other vowels are ——— and ———. Now that I have looked at the word carefully I am going to be very ——— in spelling it. I am also going to be ——— in pronouncing it. I am going to remember that the vowel in the first syllable and in the last syllable is an ———. I am not going to pronounce those syllables as if the vowel were *e* instead of ———. I am going to be very ——— about both spelling and pronouncing this word. I want it to be correct in every ———.

This device is used by providing the pupil who needs instruction with a printed or typewritten copy. The pupil is required to fill in the blank spaces correctly. This is repeated until the correct associations are fixed.

Devices for improving spelling. The following device serves to direct the pupil to see his errors in a wholesome way. It has yielded very gratifying results in the Training School of the Kansas State Normal:¹

When the spelling sentences or lists have been written each pupil is required (1) to mark each word, the spelling of which he doubts; (2) as far as possible he is encouraged to test the validity of his doubts by known means outside of the dictionary, finally checking up all doubted words by using the dictionary; and (3) he then writes all of the misspelled words, which he has thus detected, correctly spelled in separate lists; (4) at this point the pupils' papers are exchanged, the teacher spelling all words and the pupils marking those found to be misspelled on the papers; and finally (5) when the papers are returned to their owners the additional misspelled words discovered should be added to their individual lists.

The pupil's spelling is scored by the teacher on the basis of the correctness of his doubts as well as upon the number of

¹ Lull, Herbert G., "A Plan for Developing a Spelling Consciousness"; in *Elementary School Journal*, vol. 17, p. 355.

words spelled correctly. In the absence of a scientific determination of the relative significance of *spelling of words correctly* and *doubting correctly* the same value is assigned to each. The pupils are scored both for doubting words spelled correctly, and for not doubting words spelled incorrectly.

Making associations automatic. Getting the pupil to spell a word correctly is only the first step. There must be attentive repetitions of the correct associations until they have become automatic. In this respect spelling is similar to arithmetic. In the teaching of the operations of arithmetic drill occupies a prominent place, but in the case of spelling our teaching has been confined primarily to testing pupils. Requiring pupils to write each misspelled word ten or twenty times is an effort to provide practice. Such practice is unsatisfactory. After the first writing of the word the pupil probably copies. Hence the repetitions are not attentive.

Practice upon words which are misspelled by a majority of the pupils can be secured by having them recur in the spelling lesson from day to day. This plan provides the same drill for all pupils regardless of whether they misspell the word or not. In this respect it is unsatisfactory.

Courtis's spelling practice tests. In order to provide each pupil with the practice he needs Courtis has devised a series of practice tests in spelling similar to those for arithmetic. A lesson of the practice tests consists of a story with the words to be spelled printed in heavy-faced type. The pupil is directed to study these words. On the reverse side the story is printed with the spelling words omitted. A specimen lesson follows.

DETROIT PUBLIC SCHOOLS

PRACTICE TESTS IN SPELLING

Lesson No. 4. — A Trip to a Great City

On the **Sixth** of December we left the **Green Mountains** of **Vermont** to visit an uncle, who had just returned from a long voyage.

He had acquired immense wealth by deals in **leather** and **turpentine**. We were glad to arrive in **Chicago** on the **eleventh** for we had often dreamed of this visit. But how **deceiving** are dreams. Instead of being met at the depot as we expected, we found we did not know a single man or woman in all the great crowd in the station that **evening**.

We had to **argue** with ourselves to try to **understand** how anyone could **disappoint** us on such an **important** occasion.

Instructions: Read the paragraphs above and study the spelling of the words printed in heavy type until you can fill in all the blanks on the other side of this sheet correctly in four minutes.

DETROIT PUBLIC SCHOOLS

PRACTICE TESTS IN SPELLING

Lesson No. 4. — A Trip to a Great City

On the	of December we left the	1
Green	of	2
.....	to visit an uncle, who had just re-	3
turned from a long		4
He had acquired immense wealth by deals in		5
.....	and	6
We were glad to arrive in		7
on the	for we had often	8
dreamed of this visit. But how		
..... are dreams. Instead of being met at		9
the depot as we expected, we found we did not		
know a single man or woman in all the great		
crowd in the station that.....		10
We had to	with ourselves	11
to try to.....	how anyone	12
could	us on such an	13
.....	occasion.	14

Scores..... Number Tried..... Number Right.....
 Name..... Grade..... Room.....

Besides providing each pupil with the practice which he needs, and thus for individual progress, the tests have the added advantage of having each word appear in an appropriate context. A definite time is allowed, and this has been chosen so that a pupil must be able to spell the words automatically when he does the test satisfactorily.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. In what respects does the problem of measuring ability in spelling differ (1) from the problem of measuring ability in arithmetic and (2) from the problem of measuring ability in reading?
2. Is there the same need for investigations to determine the minimum essentials of arithmetic and reading as there was in the field of spelling? Why?
3. Measure the spelling ability of the pupils of your class by means of a timed sentence test and then dictate the test words as separate words. Compare the two sets of scores.
4. Teachers frequently tell with pride that all but two or three of their pupils make a "grade of 100" on a certain test. Should the fact be a cause for a feeling of satisfaction? Were the pupils really tested?
5. Dictate the words for the next spelling lesson before the pupils have studied them. Have each pupil make a list of the words which he misspells and also of the particular misspellings which he has used. Direct the pupils to base their study upon these lists.
6. How could you determine whether the method suggested in question 5 is a good one?
7. Construct a series of "timed sentence spelling tests" for the elementary school, using suitable words from the Ayres Scale.
8. How do the scores obtained by using Starch's Tests differ in meaning from the scores obtained by using a test made from the Ayres Scale?
9. Why does a test of easy words fail to give a measure of spelling ability?
10. Why must the relative difficulty of the words of a test be known if accurate measures are desired?
11. Make a study of the ways in which your pupils misspell. Also ascertain the causes for these misspellings.
12. How can you use this information in making your teaching of spelling more effective?

SELECTED BIBLIOGRAPHY

- Ashbaugh, Ernest J. *Iowa Spelling Scale for Grades II, III, and IV*. University of Iowa Extension Division, Bulletin no. 53, First Series no. 34. (Iowa City: University of Iowa, 1919. 20 pp.)

238 EDUCATIONAL TESTS AND MEASUREMENTS

- Ashbaugh, Ernest J. *Iowa Spelling Scale for Grades IV, V, and VI*. University of Iowa Extension Division, Bulletin no. 54, First Series no. 35. (Iowa City: University of Iowa, 1919. 20 pp.)
- Ashbaugh, Ernest J. *Iowa Spelling Scale for Grades VI, VII, and VIII*. University of Iowa Extension, Bulletin no. 55, First Series no. 36. (Iowa City: University of Iowa, 1919. 20 pp.)
- Ashbaugh, Ernest J. *The Iowa Spelling Scales, their derivation, uses and limitations*. Journal of Educational Research monograph no. 3. (Bloomington, Illinois: Public School Publishing Company, 1922. 144 pp.)
- Ashbaugh, Ernest J. "Variability of Children in Spelling"; in *School and Society*, vol. 9, pp. 93-98. (January 18, 1919.)
- Ashbaugh, E. J., and Horn, E. "Necessity of Teaching Derived Forms in Spelling"; in *Journal of Educational Psychology*, vol. 10, p. 143. (March, 1919.)
- Ayres, L. P. *A Measuring Scale for Ability in Spelling*. (New York: The Division of Education of the Russell Sage Foundation.)
- Ayres, L. P. *The Spelling Vocabularies of Personal and Business Letters*. (New York: Russell Sage Foundation.)
- Briggs, Thomas H., and Bamberger, Florence E. "The Validity of the Ayres Spelling Scale"; in *School and Society*, vol. 6, pp. 538-40. (November 3, 1917.)
- Gates, Arthur I. "A Study of Reading and Spelling with Special Reference to Disability"; in *Journal of Educational Research*, vol. 6, pp. 12-24. (June, 1922.)
- Hawley, W. E., and Gallup, Jackson. "The 'List' versus the 'Sentence' Method of Teaching Spelling"; in *Journal of Educational Research*, vol. 5, pp. 306-10. (April, 1922.)
- Hill, David Spence. "Standardized Illustrative Sentences for the Springfield Spelling Test"; in *Journal of Educational Psychology*, vol. 10, pp. 285-90. (May, 1919.)
- Houser, J. David. "The Relation of Spelling Ability to General Intelligence and to Meaning Vocabulary"; in *Elementary School Journal*, vol. 16, pp. 190-99. (December, 1915.)
- Jones, N. Franklin. *Concrete Examination of the Material of English Spelling*. University of South Dakota. (Vermilion: University of South Dakota, 1913.)
- Kallom, Arthur W. "Some Causes of Mis-Spellings"; in *Journal of Educational Psychology*, vol. 8, pp. 391-406. (September, 1917.)
- Morton, R. L. "The Validity of Timed-Sentence and Column Tests in Spelling"; in *Journal of Educational Research*, vol. 5, pp. 444-47. (May, 1922.)
- Otis, Arthur S. "The Reliability of Spelling Scales, Involving a 'Derivation Formula' for Correlation"; in *School and Society*, vol. 4, pp. 676-83, 716-22, 750-56, 793-96. (October 28, November 4, 11, 18, 1916.)
- Rice, J. M. "The Futility of the Spelling Grind"; in *The Forum*, vol. 23, pp. 163-72, 409-19. (April and June, 1897.)

- Scofield, F. A. "Difficulty of Ayres Spelling Scale as Shown by the Spelling of 560 High-School Students"; in *School and Society*, vol. 4, pp. 339-40. (August 26, 1916.)
- Sears, J. B. *Spelling Efficiency in the Oakland School*. A Report of the Oakland Spelling Investigation of October, 1914.
- Starch, Daniel. "The Measurement of Efficiency in Spelling, and the Overlapping of Grades in Combined Measurements of Reading, Writing and Spelling"; in *Journal of Educational Psychology*, vol. 6, pp. 167-86 (March, 1915.)
- Thorndike, Edward L. "Means of Measuring School Achievements in Spelling"; in *Educational Administration and Supervision*, vol. 1, pp. 306-12. (May, 1915.)
- Tidyman, W. F. "A Critical Study of Rice's Investigation of Spelling Efficiency"; in *Pedagogical Seminary*, vol. 22, pp. 391-400. (September, 1915.)
- Tidyman, W. F. *The Teaching of Spelling*. (Yonkers: World Book Company, 1919.) See pp. 163-78 for extensive bibliography.
- "Sixteen Spelling Scales Standardized in Sentences for Secondary Schools"; in *Teachers College Record*, vol. 21, pp. 337-91. (September, 1920.)

CHAPTER VI

ENGLISH

IN this chapter we shall consider measuring instruments in language, grammar, composition, and literature. These subjects are taught in both the elementary school and the high school, and most of the tests described may be used in both divisions of our school system.

I. THE PROBLEM OF MEASUREMENT IN ENGLISH

The nature of ability in language. Language functions in the communication of those things which are commonly called ideas and feelings by means of words. The choice and arrangement of the words give language its form. In written language, spelling and handwriting contribute additional elements of form. The ideas and feelings which language communicates may be described as its content.

The rules of grammar definitely prescribe many items of form. For example, a verb must agree with its subject in number and person; pronouns are inflected for person, case, gender, and number; verbs are inflected for mode, tense, person, and number; certain words must be capitalized. A pupil's control of those items of language form which are definite and which occur frequently should be reduced to the plane of habit or automatic functioning, so that his attention may be focused upon the content which he is attempting to express. When doubt arises concerning the form of language, the rules of grammar furnish a means of determining the correct form. Hence pupils should know the more important rules of grammar and be able to apply them in determining the correct language form. For the

fields of language and grammar the problem of measurement is largely one of measuring specific habits, and in this respect it is similar to the problem of measurement in arithmetic and reading.

Nature of ability in composition. Rhetoric treats of the choice of words and the structure of sentences and paragraphs, but it does not prescribe definite objective standards for them. The quality of these features of form is determined by the effect of the language upon the reader, and this effect is not the same for all readers. However, rhetoric does furnish certain general principles which are useful to the pupil in guiding his construction of a form which will attain his purpose. The content of language is subtle and is not objective, except as it is given a form. It depends upon the vividness and the organization of ideas, and upon the wealth of associations which give the central ideas their setting. These features of content are expressed through the choice of words and the structure of sentences and paragraphs. In this way content and form are so intimately connected that aside from the features of form which are specified by the rules of grammar, any attempt to measure one is made difficult by the presence of the other.

Ability in composition is an unanalyzed composite of these two elements. Since its quality is dependent upon the reader's estimate of the composition, certain items of form may constitute the determining factors in certain cases, but the estimate of other compositions by the same person or by different persons may be based largely upon content. This makes the problem of measuring ability in composition a difficult one in certain respects.

Nature of ability in literature. Ability in literature includes both comprehension and appreciation or enjoyment of the material read. The comprehension has already been considered in the chapter on reading. Appreciation or en-

joyment is more complex and subtle. As in the case of comprehension there are no observable objective manifestations of appreciation which can be used as an index of this trait. Furthermore, no plan has yet been devised for securing a supplementary performance which is satisfactory for measurement purposes.

II. LANGUAGE AND GRAMMAR TESTS

1. *The Charters Diagnostic Language Tests*

General structure. The Charters Diagnostic Language Tests consist of sentences, most of which involve some language error. In the case of such sentences the pupil is to write the correct word or words on a dotted line immediately below the printed sentence. For example:

May Inez and me go?

.....

"I" is to be written on the dotted line to indicate that the sentence is incorrect and also to give the correct form. There are four tests, Pronouns, Verbs, Miscellaneous A, and Miscellaneous B. Each test consists of forty exercises, one or two of which are correct. These tests may be used in Grades III to XII. Two forms of each test are available. Ample time is allowed for the pupils to finish all of the exercises.

The method of determining what language errors to include in the tests is a significant feature of their construction. From the written work of school children and from their oral language a rather complete list of the language errors which they make was obtained. This group of tests includes all the types of language errors which were found to occur at all frequently. Thus the Pronoun Test yields a "complete" measure of language ability in the field of pronouns. Similar statements can be made for the other tests. This charac-

teristic is significant when the tests are used to diagnose language difficulties. By using this series of tests one may be certain that he has "completely" canvassed the field of language and is not dealing with only a sample of the pupil's language ability.

The titles of these tests indicate that their function is diagnostic, but they also have a general or survey function. The particular function which is fulfilled depends upon the form in which the performance of the pupils is described. If a tabulation is made similar to that shown for spelling on page 227, a diagnostic function is fulfilled. If each pupil is given a score, a general or survey function is fulfilled. This dependence of the function of the test upon the plan of describing performance occurs in the case of other tests, but it is conspicuous here because Charters has used the word "diagnostic" in the title of his tests.

In scoring the test papers one credit is given for each exercise done correctly. Although some of the exercises are more difficult than others, the author decided that dropping the values of the exercises which had been calculated would not introduce serious errors in the scores. This is especially true when the tests are used for diagnostic purposes which is their announced function.

Limitations. The Charters Diagnostic Language Tests are intended to measure the ability of school children to use correct language in their oral and written expression. In the tests they are asked to correct sentences which involve language errors. One naturally questions whether a pupil's score on these tests will be a truthful index of the grammatical accuracy of his oral and written language. The tests will show whether he knows the correct forms, but it is generally recognized that "knowing" does not insure "doing," particularly in the field of language. Since no experimental evidence has been presented on this phase of the validity of

these tests, it is necessary to exercise caution in interpreting the scores which they yield.

Although two forms of each of these tests are available, no study of their reliability or of their equivalence has been reported.

Grade norms. The grade norms in Table XXIV are based upon scores obtained from Form 1. Since the degree of equivalence of the two forms is not known, caution should

TABLE XXIV. GRADE NORMS FOR CHARTERS'S DIAGNOSTIC LANGUAGE TESTS — MARCH TESTING

	GRADES					
	III	IV	V	VI	VII	VIII
*MISCELLANEOUS A						
Number of Pupils...	386	669	668	845	758	494
25-percentile.....	4.0	5.8	8.1	11.8	14.0	16.6
Median.....	6.7	9.3	11.6	16.5	18.9	22.3
75-percentile.....	13.3	13.6	16.0	21.7	24.4	27.1
†MISCELLANEOUS B						
Number of Pupils...	230	430	307	475	412	294
25-percentile.....	3.0	10.6	15.7	19.8	23.5	28.7
Median.....	7.9	17.8	22.0	27.3	29.4	32.0
75-percentile.....	14.8	24.5	27.6	32.4	33.7	36.8
**VERBS						
Number of pupils...	365	403	373	478	539	638
25-percentile.....	7.3	12.9	17.2	19.0	22.7	28.6
Median.....	12.6	17.7	22.6	24.3	27.7	32.8
75-percentile.....	18.8	22.7	28.4	29.3	31.9	36.1
PRONOUNS						
Number of pupils...	787	864	895	1344	1566	1253
25-percentile.....	8.9	11.1	14.2	17.0	19.6	23.1
Median.....	13.6	15.1	18.5	21.4	24.5	29.0
75-percentile.....	19.8	20.3	22.6	25.7	29.5	34.0

* Formerly Miscellaneous.

† Formerly Verbs B.

** Formerly Verbs A.

be exercised in interpreting scores yielded by Form 2. However, as we have pointed out in other places, when the tests are used for diagnosis, the accuracy of the norms is less vital.

2. Charters Diagnostic Language and Grammar Tests

These tests include those just described, except Miscellaneous B, plus exercises intended to measure a pupil's knowledge of the important rules of grammar. After he does a language exercise, he is asked to indicate the rule on which the correction is based. In the revised form of the "language and grammar tests," the pupil is presented with a numbered list of grammatical rules. To indicate the rule on which a given correction is based, he has only to write the number of that rule after the correction. This change makes the scoring of the test papers highly objective. There are three tests, Pronouns, Verbs, and Miscellaneous. They are intended to be used in Grades VII to XII.

3. Briggs English Form Test

This test is intended to measure ability in seven items of language form: (1) the initial capital; (2) the terminal period; (3) the terminal interrogation point; (4) the capital for a proper noun or adjective; (5) the detection and correction of a run-on sentence; (6) the apostrophe of possession; and (7) the comma before *but* coördinating the members of a compound sentence. In the test the pupil is presented with sentences which are correct except for one or more of these items. The pupil is directed to "Read over each group of words (sentences) so as to get their meaning. Then put in the proper places capital letters, apostrophes, necessary commas, periods and question marks." The seven items of form are incorporated in a cycle of four sentences. This cycle occurs five times so that the pupil has five opportunities to apply his ability on each item of form. The exercises

are intended to increase gradually in difficulty and ample time is allowed for all pupils to finish. There are two forms, Alpha and Beta. The number of items of form correctly inserted in the exercises is taken as the pupil's score.

The author studied the validity of this test by having the pupils write the sentences from dictation as well as take the test. Although the correlation between the two sets of scores is not perfect (.79), he concludes that "it is sufficiently close to warrant a high degree of confidence" in his test as an instrument for measuring the ability to use these items of language form. He, however, states that the scores on the test are considerably lower than when the sentences were dictated. Thus, when considered as an instrument for measuring absolute ability, this test is lacking in validity. Pupils may be expected to make relatively fewer errors in their regular written work than they make on this test. In addition to this limitation one should bear in mind that the coefficient of reliability given is .76.

4. Kirby Grammar Test

This test is designed to measure a student's knowledge of "correct English usage" and his ability to "select the rule or principle in accordance with which a usage is correct." In the left-hand column a list of sentences is given. In each of these sentences two forms are printed in parentheses, one correct and the other incorrect. In the left-hand column a list of rules or principles is given. The pupil is asked to mark out the incorrect form and to indicate the rule which tells why this form is incorrect. There are five sections each including from eight to ten exercises. The items of language form on which the exercises are based were determined by study of the errors which school children make. In this respect the test is similar to those devised by Charters. The structure of this test differs from that of the Charters Lan-

guage and Grammar Tests in one important respect. Charters requires the pupil to detect the incorrect form and to correct it. Kirby merely asks the pupil to choose between two forms which are presented. Thus guessing may be a potent factor in determining a pupil's performance.

5. *Pressey Diagnostic Tests in English Composition*

This group of tests includes one each on vocabulary, grammar, and punctuation. In the vocabulary test the pupil is given sentences which are complete except for one word. He is asked to choose from a list of four words printed immediately after the sentence the one which best completes the meaning of the sentence. In the grammar test an exercise consists of four sentences only one of which is grammatically incorrect. The pupil is to find the "wrong" sentence in each exercise and to draw a line under it. In the punctuation test the pupil is given sentences which lack all punctuation and capitalization except the capital at the beginning and the period at the end. The pupil is asked to insert the necessary punctuation and capital letters. This battery of tests is intended to measure separately three fundamental abilities which function in the writing of compositions.

6. *Starch Grammatical Scales*

Starch¹ has devised three scales (A, B, and C) to measure a pupil's ability to use correctly certain language forms. His Grammatical Scale A consists of a series of exercises such as the following:

STEP 7

1. A fireman seldom rises above (an engineer; the position of an engineer).

¹ Starch, Daniel, "The Measurement of Achievement in English Grammar"; in *Journal of Educational Psychology*, vol. 6, pp. 615-26; also in his *Educational Measurements*, pp. 105-08.

2. The difference between summer and winter (is that; is) summer is warm and winter is cold.
3. He is happier than (me; I).
4. They are (allowed; not allowed) to go only on Saturday.

The pupil is given a printed copy of the scale and is directed thus: "Each of the following sentences gives in parenthesis two ways in which it may be stated. Cross out the one you think is incorrect or bad. If you think both are incorrect, cross both out. If you think both are correct, underline both." Pupils are given as much time as they need.

The sentences of the exercises have been chosen so that the difference in difficulty between any two successive steps of the scale is equal to the difference between any other two successive steps. A pupil's score is the highest step of the scale of which he does correctly three out of the four sentences. If a pupil fails on a given step, say the seventh, but does the ninth correctly, his score is 8. Thus, a pupil receives credit for each exercise of which he does correctly three out of the four sentences, but the sentences have been so arranged that only in a few cases will a pupil be able to do an exercise correctly after he has missed the preceding.

As tentative norms of attainment Starch gives the following scores for the use of these scales:

Grade.....	VII	VIII	IX	X	XI	XII	Freshmen
Score.....	8.0	8.3	8.6	8.9	9.2	9.5	10.3

The scale includes many different items of grammatical form and apparently these are arranged in no systematic manner. Therefore, a pupil's score can be only a general measure of his ability to use correct language forms. To secure detailed information concerning his weaknesses it would be necessary to examine his test paper.

7. *Starch Punctuation Scale*

Starch has also devised a Punctuation Scale which is similar in form to the Grammatical Scale. The exercises consist of sentences to be punctuated. The following extracts illustrate the nature of this scale.

STEP 6

1. We visited New York the largest city in America
2. Everything being ready the guard blew his horn.
3. There were blue green and red flags.
4. If you come bring my book.

STEP 12

1. When thou goest forth by day my bullet shall whistle past thee when thou liest down by night my knife is at thy throat.
2. Oh come you'd better.
3. The president bowed then Hughes began to speak.

A pupil's score is determined in the same way that it was in the case of the Grammatical Scale. In each scale certain features of the form of language have been isolated for the purpose of measurement. Tentative norms of attainment are the same as for the Grammatical Scale.

8. *Wilson Language Error Test*

The Wilson Language Error Test, Story A, consists of a short story in which there are twenty-eight language errors. The pupil is asked to "correct all the mistakes in the story." These errors are varied in type. Some refer to verb forms, others to the use of pronouns, and still others to the choice of appropriate words. The errors in the story are the ones which investigation has shown that children commonly make. Story B and Story C are similar to Story A and are intended to be used as duplicate forms. The author suggests that Story A be used at the beginning, Story B at the

middle, and Story C at the end of the year. The test is designed to be given in Grades III to XII inclusive, although it is a little difficult for pupils in the third grade. The author suggests the use of this test for diagnostic purposes. By means of it a teacher can discover the language errors which her pupils are accustomed to make and can direct her training to remedy these defects.

9. *Boston Copying Test*

The test. Copying is a phase of school work which receives little explicit attention. This probably is due to the assumption that pupils are able to copy accurately because it appears to be such a simple activity. Copying bears a relation to written expression and to other school subjects as well. Themes are usually copied before being submitted to the teacher. In solving problems in arithmetic the quantities are copied from the text. In gathering information from references copying occurs. The following test of pupils' ability to copy printed matter was prepared by a group of Boston ¹ teachers:

DIRECTIONS FOR GIVING AND SCORING THE TEST

1. Read to the pupils the directions which are printed at the head of the selection they are to copy, but give them no further help. For example, do not specify possible errors which may be made.
2. Pupils ought not to see the selection until they are ready to copy it. Hence it should be placed on the desk face down until the signal is given to begin work.
3. Every error should be checked distinctly.
4. The errors which were to be noted were as follows: In spelling, capitalization, punctuation, undotted "i's," uncrossed "t's"; in omitting words, in adding words, in wrong words used, and in misplaced words.

¹ *School Document no. 2, 1916.* (Boston Public Schools, English, *Determining a Standard in Accurate Copying.*)

DIRECTIONS TO PUPILS

Copy in ink as much of the following selection as you can copy accurately in fifteen minutes without hurrying. Accuracy is more important than speed:

LIEUTENANT OULESS

In this story a young British lieutenant, in a moment of extreme irritation, strikes a private soldier. The act is one that calls for dismissal from the Queen's service. What is the officer to do? He cannot send money to the soldier — who happens to be the redoubtable Ortheris himself — nor can he apologize to him in private. Neither can he let matters drift. Ortheris, too, has his own code of pride and honor; he too is a "servant of the Queen"; but how is the insult to be atoned for? The way out of this apparently hopeless muddle is a beautifully simple one, after all. The lieutenant invites Ortheris to go shooting with him, and when they are alone, asks him "to take off his coat." "Thank you, sir!" says Ortheris. The two men fight until Ortheris owns that he is beaten. Then the lieutenant apologizes for the original blow, and the officer and private walk back to camp devoted friends. That fight is the moral salvation of Lieutenant Oules. (Bliss Perry, *A Study of Prose Fiction*.)

Kinds of errors made. This test was given to 4494 first-year pupils in the Boston high schools in November, 1914, and therefore may be considered to measure the ability of pupils completing the eighth grade. The results are both interesting and significant. The following is quoted from the bulletin mentioned above:

The errors noted consisted of nine different kinds, and the number of each kind made in this test by 4494 pupils is shown by the following tabulation:

Spelling.....	5829
Capitalization.....	644
Omitted words.....	4077
Added words.....	606
Wrong words used.....	840
Misplaced words.....	105
Punctuation.....	5876
Undotted "i's".....	8794
Uncrossed "t's".....	606
Total.....	27,377
Average errors per pupil.....	6.09

Misspelled words. The test consisted of 170 words, 105 of them different words. It is a notable fact that every word was misspelled by somebody. It is also interesting that 92.2 per cent of the words in the test are found in Jones's *Concrete Investigation of the Material of English Spelling*.¹ In spite of the fact that these are words commonly used by children in their writing, 11.8 per cent of them were misspelled more than 100 times. This does not mean that 11.8 per cent of the children missed these words, because one pupil might have missed the same word more than once.

It is impossible to make any statement in regard to the average because many of the words occur in the selection more than once, and if misspelled by the same person each time it occurs it is counted more than one error. Some children spelled a word incorrectly in one place and correctly in another. One boy spelled "lieutenant" wrong four out of five times, and spelled it a different way each time. Then, not all the children finished the entire selection, and no record was kept of the exact number of words each wrote. However, 4494 pupils taking the test made 5829 errors in spelling alone, the number of errors for each word varying from 1 to 1045.

Undotted "i's" and uncrossed t's. The errors made by leaving the "i's" undotted and the "t's" uncrossed comprise about one third of the entire number of errors and are largely important because of their value to legibility, as pointed out by Ayres. In connection with these errors, it is very noticeable that most of them were confined to comparatively few pupils. If a child showed a tendency to dot his "i's" and cross his "t's" in the first few lines, the chances were that that individual would have but few errors. On the other hand, if the child made many errors in the first part of the paper, there were many throughout the copying. One boy went through the entire paper without dotting an "i." Many others dotted only a small part of them.

This same test was given in Kansas City, Missouri, to the pupils in the seventh grade and in the first year of the high school. (Kansas City has only seven grades below the high school.) The average errors per pupils was 8.04 in the seventh grade,² and 6.83 in the first year of high school.

¹ See page 228.

² Corresponds to the usual eighth grade.

III. ENGLISH COMPOSITION

Ability in English composition is measured by requiring pupils to write compositions which are then described in terms of a composition scale, similar to the handwriting scales described in Chapter IV. Thus the consideration of the measurement of ability in English composition involves the directions for securing the compositions as well as the scales which are used to describe them.

1. Directions for securing compositions

Points to be covered in the directions. The purpose to be accomplished by these directions is to secure compositions which will be equally representative of the ability of all pupils to write compositions under known conditions. Thus there should be specifications with reference to (1) type of subject, (2) time to be used, and (3) rewriting for making corrections.

Hudelson experimented with a variety of types of subjects and concluded that pupils write their best compositions on the topic, "How I Learned a Lesson," and the composition most typical of their ability when reproducing a story read to them. He also states that "contrary to expectations, pupils who are allowed to choose their own theme subjects produce results highly indicative of their average composition achievement." In his investigation fifteen minutes were allowed for the writing.

It is generally agreed that the topic on which the pupils write should be similar to that on which the scale compositions are written. Hence a teacher should be guided in his choice of topics by the scale which he uses in rating the compositions obtained.

Directions for using the Willing Scale. The following directions accompany the Willing Scale. They illustrate

the details which should be specified if the compositions are to be written under controlled conditions.

The teacher should write on the blackboard these topics:¹

An Exciting Experience

A Storm

An Accident

An Errand at Night

A Wonderful Story

An Unexpected Meeting

In the Woods

In the Mountains

On the Ice

On the Water

A Runaway

The teacher should then say to the pupils: "I want you to write me a story. It is to be a story about some exciting experience that you have had, about something very interesting that has happened to you. If nothing of the sort has ever happened to you, then tell me of an exciting experience some one you know has had. You may even make up a story of this kind, if you have to, though I believe you will do better, on the whole, with a real one. I am going to give you about twenty minutes in which to write. You are to write on both sides of the paper, to do all the work yourselves, and to ask no questions at all after you begin. You may make whatever corrections you wish between the lines. There will be no time to rewrite your story.

"I have written the general subject on the board together with some suggestions. You do not have to write on any of these topics unless you want to; they are merely to help out in case you cannot think of an exciting experience yourself. You may begin now as soon as you wish."

Allow opportunity for asking questions and make an effort to put the children at ease. Allow full twenty minutes for the actual writing. At the end of this time say to the pupils: "You are to have four or five minutes in which to finish your stories, make corrections and count the number of words written. Write this

¹ It is probably better to furnish each pupil with a printed list of the topics.

number at the end of your story. Also write your name and school grade." At the end of five minutes collect the papers.

Directions for using the Hillegas Scale. The Hillegas Scale has been used in the surveys of Butte, Montana, and Salt Lake City, Utah. At Salt Lake City the following directions were followed for securing compositions written by pupils:

1. Each teacher is requested to ask her children to write a composition for her on the following theme:

"Suppose that you have twenty dollars, which you have been given to spend. You have five friends, and you decide to spend it in such a manner as will give the most pleasure to each. Tell what you would do or buy for each friend. The amount spent for each friend need not be the same, but the total for the five must be twenty dollars."

2. The composition should be written with pen and ink on the regular writing paper.

3. After the children are ready for writing, read the subject to them, give them a minute or two to ask any questions, and as soon as you are sure that the children understand what they are to do, start them at writing.

4. When the children have finished collect the papers, fasten those for each class together with a clip, and send to the office of the school principal.

2. *Composition scales*

General structure. A composition scale consists of a series of suitable compositions arranged in ascending order of merit. The numerical description of the merit of a scale composition is taken as its scale value.¹ Existing composition scales differ in respect to the scope of the "merit" which they are intended to measure. Such scales as the Hillegas, Nassau County Supplement, and Hudelson yield

¹ For a description of the method of scale construction, see Monroe, Walter S., *An Introduction to the Theory of Educational Measurements*, chap. vi. (Houghton Mifflin Company, Boston, 1923.)

measures of "general merit." The quality "general merit" has not been objectively defined but it may be thought of as an unanalyzed composite of the various qualities whose presence contributes to the excellence of a composition. These qualities include story value, diction, unity, clearness, sentence structure, grammatical accuracy, punctuation, spelling, etc. The Willing Scale and Van Wagenen Scale are intended to yield separate measures of certain of the qualities. In other words, they represent attempts to secure diagnostic measurements of ability in English composition.

The composition scales designed to measure general merit differ in respect to (1) the character of the scale compositions (form of discourse, topic, conditions of writing, etc.), (2) the number of degrees of quality represented in the scale, and (3) the values of the compositions chosen for the scale. Our best scales consist of compositions which are similar in general character and which were written under much the same conditions as those to be followed in securing compositions from pupils when using the scale. The values of the scale compositions should be accurately determined and the steps of the scale should be equal. It is desirable that the lower end of the scale should approximate zero merit and that the best compositions should approach perfection. It appears that except for highly trained persons, a scale of ten steps is as precise an instrument as one can profitably use.

The Hillegas Scale. The Hillegas Composition Scale was constructed in 1912. Since then better scales have been made but it is of interest to students of educational measurements because it was the first composition scale and also because several of the more recent scales are essentially revisions of it. The Hillegas Composition Scale consists of ten compositions ranging from an artificial production whose scale value is zero to the tenth composition whose scale value is 9.3. Three of the ten compositions are artificial produc-

tions, five were written by high-school pupils, and the remaining two by college freshmen. No two were written on the same topic and they vary greatly in length and type. Each degree of merit is represented by only one composition. In the Thorndike Extension of the Hillegas Scale only a few of the compositions of the original scale have been used and several compositions are given for each degree of merit in the middle of the scale. Twenty-nine compositions represent fifteen degrees of merit within approximately the same range as the original scale. This makes a more finely divided scale than the original one. The unit in terms of which "general merit" is measured by means of this scale is that difference of "general merit" which was noticed by exactly seventy-five per cent of the judges to whom the scale compositions were submitted. If we take two compositions, A and B, and if seventy-five per cent of the judges say A is better than B while the remaining twenty-five per cent say that B is better than A, then the difference in "general merit" between these two compositions is one unit. This unit is relatively large. The increase in the norms from grade to grade is about one half of a unit.

Nassau County Supplement to the Hillegas Scale. In the course of a survey of the elementary schools of Nassau County, Long Island, made during 1916, a composition scale known as the Nassau County Supplement was constructed. It may be thought of as a revision of the Hillegas Scale. With the exception of three compositions at the upper end of the scale, the scale compositions were written by elementary school pupils on the topic "What I should like to do next Saturday." The scale represents ten steps of merit and is printed on a single sheet in a form convenient for use. It is one of the best scales for measuring the general merit of English composition. The steps of the scale are slightly irregular and it has also been criticized because of the shortness of the scale compositions.

Willing Scale. Willing used compositions written by pupils in grades four to eight on the topic, "An Exciting Experience." Several particularly exciting experiences were suggested, and twenty minutes were allowed for writing. In determining the compositions to be used for the scale, "all errors in spelling, punctuation, capitalization, and grammar were counted and corrected." The relative merit of the corrected compositions was determined and those compositions were selected for the scale which had the same rank in "story value" and frequency of errors. In the printed scale the compositions include the errors and in addition to the scale value, the number of errors per hundred words is stated. The scale is intended to yield separate measures of "story value" and "form value." The measure of "form value" is to be obtained by counting the number of errors in "spelling, punctuation, capitalization, and grammar." Thus this measure approaches objectivity. The scale has been criticized on the basis of the statistical procedure which Willing used in constructing it. However, it has been widely used and is one of the most helpful of the existing composition scales.

Hudelson Scales. The Hudelson English Composition Scale is essentially a revision of the Hillegas Scale. The most significant feature is that the steps of the scale are .5 of a unit instead of one unit as is the case in the Hillegas Scale and the Nassau County Supplement. This refinement leads one to expect that more accurate measures would be obtained, but Hudelson states that "the steps are too finely divided for all save very highly trained judges dealing with small groups or individual themes." The scale is printed in pamphlet form which makes it inconvenient to use.

Hudelson has recently devised two other scales: "Maximal Composition Ability Scale" and "Typical Composition Ability Scale." The first consists of compositions written

on the topic "How I Learned a Lesson." The second consists of reproductions of a story. Hudelson presents data to show that in general pupils will write their best compositions on the first topic and typical or average compositions on the second. Both of these scales may also be considered revisions of the Hillegas Scale. The steps of the scale are approximately one unit.

Van Wageningen English Composition Scales. These scales represent an attempt to construct a group of instruments which would yield diagnostic measures of ability in English composition. There is a separate scale for narration, description, and exposition. The compositions of each scale have been evaluated with respect to three qualities, "thought content," "structure," and "mechanics." It is intended that by means of one of these scales a pupil composition will be rated three times, once for each of the three qualities. The rating for each of these qualities is to be independent of the other two. For example, in rating for "thought content" both of the other two qualities are to be neglected. This makes the use of the scale rather complex, and teachers report that they are unable to use it satisfactorily.

Lewis Composition Scales. Lewis has devised separate scales for the following types of writing: (1) simple order letters; (2) letters of application; (3) simple narrative social letters; (4) expository social letters; and (5) narratives on the topic, "One of my most interesting experiences." The first four of these scales are unique and should prove helpful in measuring the types of writing which they represent. These scales are now published on separate sheets instead of as a monograph.

Courtis Standard Research Test in Composition. Courtis has recently published a composition scale which is unique in its form and the method of its use. Each scale composi-

tion is a facsimile reproduction of the original hand-written composition and is printed on a single page of regular theme paper. There is no mark on the front of this sheet to indicate its scale value. This is printed on the back. Thus instead of a composition scale printed on a single large sheet or in pamphlet form we have one on separate sheets of theme paper. In using this scale Courtis directs that the pupils be required to use the same kind of theme paper. Instead of "matching" the pupil compositions with the scale compositions, the latter are mixed with the former. All compositions are read and sorted into five piles representing five degrees of merit. Then the papers are reread and arranged in order of merit. After this has been completed the scale compositions are located. If the scorer has arranged them among the other compositions in the order of their scale values, it is assumed that the scoring has been done satisfactorily. If the scale compositions have been placed in some other order, the scorer's "judgment of the general value of compositions differs from that of most teachers," and the scoring is considered to be unsatisfactory. The compositions are mixed thoroughly and sorted again. Although we do not have any experimental evidence it is likely that this plan of measuring ability in written composition will tend to increase its reliability.

3. *Using composition scales*

Plan of use. The general plan for using composition scales is similar to that described for handwriting scales in Chapter IV. A pupil composition is to be "matched" with the scale composition which it most nearly resembles in "general merit" or the particular characteristic being measured. The value of the scale composition which it most nearly resembles is taken as the measure of the pupil's composition. In comparing a pupil composition

with the scale composition for "general merit" one should avoid being unduly influenced by particular features of the composition, such as errors in spelling, grammatical errors, sentence structure, etc. When rating compositions with reference to a particular characteristic such as "story value" or "structure" all other characteristics should be neglected.

One should have the composition scale spread out before him so that all steps are in view. When a scale is printed in pamphlet form this is not possible. When a pupil's composition appears to fall between two scale compositions it is not advisable to assign intermediate values. A good plan to follow in rating compositions, especially for teachers who are inexperienced in the use of composition scales is to sort the compositions in piles corresponding to the steps of the scale. After this has been done, each pile should be taken separately and the compositions compared with each other and with the scale. Then any readjustments which seem desirable should be made before values are assigned to the compositions.

Two teachers will differ in the scores which they give to a large per cent of a group of compositions. Experience in using composition scales has shown that a few hours of practice in rating compositions will greatly reduce the subjectivity of the process. The plan for practice is similar to that recommended for handwriting scales on page 183. Any group of compositions may be used, but it will be better to use a set whose true scores are known. Thorndike has prepared a set of "150 specimens arranged for use in psychological and educational experiments."

Grade norms. The scales by Hillegas, Thorndike, Trabue (Nassau), Hudelson and Lewis are expressed in terms of the same unit and from the same zero point. Hence the same grade norms may be used with all of these scales. Hudel-

son¹ has recently published the following "National Standards." These are based upon scores reported from schools in different sections of the United States. He also gives grade norms for a number of cities.

"NATIONAL STANDARDS" FOR JANUARY

GRADE	NORM
IV	3.0
V	3.6
VI	4.2
VII	4.7
VIII	5.3
IX	5.5
X	5.9
XI	6.3
XII	6.7

Reliability of measures of ability in English composition. The reliability of a single measurement of ability in English composition depends upon the variability of successive performances of the pupils as well as upon the objectivity of the rating of the compositions secured. Many pupils will write better compositions on one topic than on another. The performance of all pupils will be determined by their past experience and present frame of mind. In the case of tests whose scoring is highly objective we find surprisingly large differences between two applications of the test. It is reasonable to expect that the variations in performance would be even greater in the case of written compositions. In addition, the rating of compositions by means of any of the scales described on pages 255-259, is far from perfectly objective. Trained judges will differ in the scores assigned

¹ Hudelson, Earl, "English Composition: Its Aims, Methods, and Measurement"; in *Twenty-Second Yearbook of the National Society for the Study of Education*, part I, p. 152. (Bloomington, Illinois: Public School Publishing Company, 1923.)

to many compositions. In one investigation eighty-six compositions were scored independently by two persons using the Willing Scale. The coefficient of correlation between the two sets of scores was .86. In addition there was a difference of 6.7 between the averages of the two sets of scores. This indicates the presence of a relatively large constant error.

The accuracy of measurements made by using the Hillegas Scale has been investigated by having the same compositions rated by a group of teachers, first, by the usual method and second, by using the scale. By means of a number of such investigations the conclusion has been reached that "the variability is somewhat greater with the scale than without it."¹ However, in the investigations reported the teachers using the scale were untrained in its use. Furthermore, as in the case of handwriting practically no attention has been given to determining the best methods of using a composition scale. Because of these two facts the conclusions reached in the studies just referred to must be qualified. Thorndike has asserted that errors in using the scale will diminish with practice² and with sufficient practice they will be smaller than the errors now made by teachers in grading paragraph writing for general merit. Trabue states that "In spite of all criticisms of and objections to the Hillegas Scale, the fact remains that it is one of the most useful measuring instruments in the whole field of education."³

Reliability when corrected for "mechanical errors." An interesting sidelight upon the validity of the measurement of ability in written composition was given by one investiga-

¹ Kelly, F. J., *Teachers' Marks*, p. 134.

² Thorndike, E. L., "Notes on the Significance and Use of the Hillegas Scale for Measuring the Quality of English Composition"; in *The English Journal*, vol. 2, p. 551.

³ Trabue, M. R., "Supplementing the Hillegas Scale"; in *Teachers College Record*, vol. 18, p. 51.

tion.¹ The six compositions comprising the Harvard-Newton Exposition Scale were reproduced without any identifying marks. They were graded on the scale of 100 per cent by twenty-four eighth-grade teachers who were asked to follow certain typewritten directions. The six compositions were then "completely corrected so far as mechanical or measurable errors were concerned." The corrected compositions were graded by the same teachers according to the same directions.

If the "mechanical errors" of the compositions were significant factors in determining the first set of marks, the second set of marks should be conspicuously higher. However, this was not the case. For two of the compositions the average "grade" was less after the "mechanical errors" had been corrected. The individual marks show that some teachers consider form important, and that others tend to disregard it in marking a composition.

Hudelson's index of reliability. Hudelson has recently studied the reliability of the measurement of ability in written composition by having pupils write two compositions on similar topics under similar conditions and then calculating the coefficient of correlation between the pairs of scores thus obtained. The median of the independent ratings by eight trained judges was used as the score. Thus we should expect to find that the coefficients of reliability obtained in this way would be materially higher than if single ratings had been used. The coefficients of reliability obtained by Hudelson range from .69 to .82. It is estimated that if he had used single ratings the coefficients of reliability would be in the neighborhood of .40. This index of reliability takes no account of constant errors in the score due to the tendency of a scorer to assign "high" or "low" marks.

¹ Brownell, Baker, "A Test of the Ballou Scale of English Composition"; in *School and Society*, vol. 4, pp. 938-42.

Thus it is clear that in comparison with educational measurements in such fields as arithmetic and reading, our measurements of ability in English composition are very unreliable. It, however, appears that by means of composition scales we may secure more reliable measures than we are able to obtain without them. In addition norms are available for interpreting the measures.

IV. LITERATURE

The problem of measurement. The nature of the achievements in literature creates a very difficult problem of measurement. This is doubtless the reason why so few attempts have been made to construct standardized tests in this field. Van Wagenen has constructed a reading scale in the field of literature. It is merely a scaled reading test in which the pupil is asked to check those statements in a given list which contain ideas in the paragraph or can be derived from it. The paragraphs are selections from literature texts. Obviously such a test can measure only one outcome for which we strive in the teaching of literature.

Abbott and Trabue have devised a scale to measure ability to judge poetry. The scale consists of thirteen exercises each of which contains four versions of a short poem. The pupil is asked to select the version which he likes best as poetry and also the one which he likes least. One version is the original poem. In each of the other three changes have been made which lower its quality. In one version ("sentimental") it was attempted "to falsify the emotion by introducing silly, gushy, affected, or otherwise insincere feeling; in another version (the 'prosaic') to reduce the poet's imagery to a more pedestrian and commonplace level; in a third (the 'metrical') to render the movement either entirely awkward or less fine and subtle than the original." The authors found that the scale was not useful in the elementary school

and for high-school students the reliability was found to be only .440. Hence the scale has only a very limited value for the high school. One may also raise the question whether the scale measures the total outcome of our teaching of poetry.

V. EDUCATIONAL SIGNIFICANCE OF THE USE OF THESE TESTS AND SCALES

Finding specific language weaknesses. The tests in language and grammar are intended to measure certain specific language abilities. If we assume that this intended function is fulfilled, they furnish the teachers with detailed information about the language achievements of their pupils. If a pupil is weak in punctuation, the test reveals the fact. The teacher then knows that he must instruct that pupil in punctuation. Similar statements can be made with reference to the other phases of language ability measured by the available tests. The grammar tests described in this chapter measure one type of acquaintance with the rules of grammar. In interpreting the scores which they yield, one must bear in mind that the ability measured may not be the total outcome which is desired from instruction in grammar.

The composition scales fulfill the same function in the field of language as the handwriting scales of Ayres, Thorndike, Starch and others do in that field. They are instruments for general measurement. By means of them a teacher can obtain a measure of certain language abilities of his pupils in terms of fixed units which he may compare with established standards or with similar measures of other groups of pupils.

Remedying the situation revealed. When a teacher learns the specific language weaknesses of his pupils he is then in position to apply more intelligently his stock of methods and devices of instruction. In language as in the case of the

other subjects, the teacher must instruct individual pupils who are grouped together rather than groups of pupils. Furthermore, each pupil should receive the instruction which he needs to correct his language errors.

If pupils are weak in a language ability, such as punctuation, the laws of habit-formation apply. After being sure that he understands the function of the punctuation marks, a pupil must have practice in punctuating his own writing. This probably is not sufficient. Exercises for practice can be constructed by taking appropriate material and reproducing it without the punctuation marks.

Until a teacher recognizes definite and specific ends to be attained there is certain to be a large degree of dissipation of his efforts. Perhaps one reason why language instruction so often does not produce satisfactory results is that it is not directed toward the engendering of definite abilities. In teaching spelling, teachers have kept a record of pupils' errors and have emphasized these words in their teaching. In our consideration of spelling it was urged that teachers first ascertain what words their pupils were unable to spell correctly. This plan may be adapted to the teaching of other aspects of language. The teacher should ascertain the pupils' grammatical errors, and then equip them with the rules of grammar which are needed to correct them.

Perhaps the scales and tests described in this chapter will have fulfilled their most important function if they cause teachers to analyze and define "language ability" in more specific terms. It is believed that their use will tend to produce this result. Analysis of "language ability" and specific definition of the elements are greatly needed. Upon the accomplishment of these two things depends the construction of more valuable measuring instruments in the language field and the scientific determination of methods and devices of instruction.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. How does the problem of measurement in language differ from the problem of measurement in arithmetic?
2. What makes the problem of measurement in language difficult?
3. How does the Hillegas Scale differ from the Willing Scale? The compositions which compose the Willing Scale were written under defined conditions and on similar topics. Does this make it a superior scale?
4. Which of the composition scales described in this chapter will be the most helpful to the teacher? Why?
5. Give the copying test to your pupils following the directions carefully. Do the results agree with your estimate of the ability of your pupils to copy?
6. Keep accurate lists of the language errors of your pupils. What are the rules which are necessary to correct these errors? Are they the rules upon which you are placing the most emphasis in your teaching?
7. Do you have definite objective standards of attainment in English composition? Can you use the tests described in this chapter to establish such standards?
8. Do you think pupils would be helped by having definite objective standards of attainment established for them?
9. What is the value of a diagnostic test in language and grammar?
10. Does Charters's method of constructing his tests insure a complete diagnosis?
11. Why is the measurement of ability to write compositions less reliable than the measurement of ability in arithmetic or silent reading?

SELECTED BIBLIOGRAPHY

- Abbott, Allan, and Trabue, M. R. "A Measure of Ability to Judge Poetry"; in *Teachers College Record*, vol. 22, pp. 101-26. (March, 1921.)
- Ashbaugh, Ernest J. "The Measurement of Language: What is Measured and its Significance"; in *Journal of Educational Research*, vol. 4, pp. 32-39. (June, 1921.)
- Ballou, Frank W. *Scales for the Measurement of English Compositions*. The Harvard-Newton Bulletins, no. 11. (Cambridge, Massachusetts: Harvard University, 1914. 93 pp.)
- Ballou, Frank W. *English — Determining a Standard in Accurate Copying*. School Document no. 2, Department of Educational Investigation and Measurement Bulletin, no. 6. (Boston: Public Schools, 1916. 25 pp.)
- Ballou, Frank W. *Harvard-Newton Composition Scales*. Harvard Bulletin, no. 2. (Cambridge, Massachusetts: Harvard University.)
- Breed, F. S., and Frostic, F. W. "A Scale for Measuring the General Merit of English Composition in the Sixth Grade"; in *Elementary School Journal*, vol. 17, pp. 307-25. (January, 1917.)

- Briggs, Thomas H. "English Composition Scales in Use"; in *Teachers College Record*, vol. 23, pp. 423-52. (November, 1922.)
- Briggs, Thomas H. "An English Form Test"; in *Teachers College Record*, vol. 22, pp. 1-11. (January, 1921.)
- Brown, M. D., and Haggerty, M. E. "The Measurement of Improvement in English Composition"; in *English Journal*, vol. 6, pp. 515-27. (October, 1917.)
- Brownell, B. "A Test of the Ballou Scale of English Composition"; in *School and Society*, vol. 4, pp. 938-42. (December 16, 1916.)
- Certain, C. C. "By What Standards are High-School Pupils Promoted in English Composition?" in *The English Journal*, vol. 10, pp. 305-15. (June, 1921.)
- Charters, W. W. "Constructing a Language and Grammar Scale"; in *Journal of Educational Research*, vol. 1, pp. 249-57. (April, 1920.)
- Charters, W. W., and Miller, Edith. *Course of Study Based upon Grammatical Errors of School Children of Kansas City, Missouri*. (University of Missouri, 1915.)
- Darsie, Marvin L. "The Reliability of Judgments Based on the Willing Composition Scale"; in *Journal of Educational Research*, vol. 5, pp. 89-90. (January, 1922.)
- Dolch, Edward William. "More Accurate Use of Composition Scales"; in *The English Journal*, vol. 11, pp. 536-44. (November, 1922.)
- Gordon, Kate. "A Class Experiment with the Hillegas Scale"; in *Journal of Educational Psychology*, vol. 9, pp. 511-13. (November, 1918.)
- Greene, Harry A. "Tests for the Measurement of Certain Phases of Linguistic Organizations in Sentences"; in *Journal of Educational Psychology*, vol. 11, pp. 517-25. (December, 1920.)
- Gunther, Charles. "My Experience with the Hillegas Scale"; in *The English Journal*, vol. 2, pp. 535-42. (November, 1919.)
- Hillegas, M. B. "A Scale for the Measurement of Quality in English Composition for Young People"; in *Teachers College Record*, vol. 12. (September, 1912.)
- Hooper, C. L. "Second-Grade Composition Scale for Second and Third Grade Teachers"; in *Chicago Schools Journal*, vol. 4, pp. 127-32. (December, 1921.)
- Hudelson, E. "Some Achievements in the Establishment of a Standard for the Measurement of English Composition in the Bloomington, Indiana Schools"; in *The English Journal*, vol. 5, pp. 590-97. (November, 1916.)
- Hudelson, E. "English Composition: Its Aims, Methods, and Measurement"; in *Twenty-Second Yearbook of the National Society for the Study of Education*, part 1. (Bloomington, Illinois: The Public School Publishing Company, 1923. 172 pp.)
- Jamison, Grace S. "A Study in Correlation of Allied English Abilities"; in *Journal of Educational Research*, vol. 6, pp. 241-53. (October, 1922.)

270 EDUCATIONAL TESTS AND MEASUREMENTS

- Johnson, F. W. "The Hillegas-Thorndike Scale for Measuring the Quality in English Composition by Young People"; in *School Review*, vol. 21, pp. 39-49. (January, 1913.)
- Jordan, R. H. "A Threefold Experiment in High-School English"; in *The English Journal*, vol. 10, pp. 560-69. (December, 1921.)
- Kayfetz, Isidore. "A Critical Study of the Hillegas Composition Scale"; in *Pedagogical Seminary*, vol. 21, pp. 449-77.
- Kirby, Thomas J. "A Grammar Test"; in *School and Society*, vol. 11, pp. 714-19. (June 12, 1920.)
- Melvin, Arthur Gordon. "A True-False Test in English Literature"; in *The English Journal*, vol. 11, pp. 491-96. (October, 1922.)
- Potter, Howard Eugene. *Abilities and Disabilities in the Use of English Found in the Written Compositions of Entering Freshmen at the University of California*. University of California, Bureau of Research in Education Study, no. 12. (Berkeley: University of California, 1922. 51 pp.) (A thesis.)
- Pressey, Sidney L. *Measurement of Progress in English in the Upper Grades*. University of Indiana Extension Division Bulletin, vol. 6, no. 12, pp. 35-45. (Bloomington: University of Indiana, 1921.)
- Pressey, Sidney L. "Scale of Attainment No. 2 — and Examination for Measurement in History, Arithmetic, and English in the Eighth Grade"; in *Journal of Educational Research*, vol. 3, pp. 359-69. (May, 1921.)
- Sackett, L. W. "Comparable Measures of Composition"; in *School and Society*, vol. 5, pp. 233-39. (February 24, 1917.)
- Starch, Daniel. "The Measurement of Achievement in English Grammar"; in *Journal of Educational Psychology*, vol. 6, pp. 615-25. (December, 1915.)
- Theisen, W. W. "Improving Teachers' Estimates of Composition"; in *School and Society*, vol. 7, pp. 143-50. (February 2, 1918.)
- Thorndike, E. L. "A Scale for Measuring the Merit of English Writing"; in *Science*, vol. 33, pp. 935-38. (June 6, 1911.)
- Thorndike, E. L. "Notes on the Significance and Use of the Hillegas Scale for Measuring the Quality of English Composition"; in *English Journal*, vol. 2, pp. 551-61. (November, 1913.)
- Trabue, M. R. "Supplementing the Hillegas Scale"; in *Teachers College Record*, vol. 18, p. 51. (January, 1917.)
- Van Wagenen, M. J., and Kelley, Frances E. "Language Abilities and their Relations to College Marks"; in *Journal of Educational Psychology*, vol. 11, pp. 459-73. (November, 1920.)
- Van Wagenen, M. J. "The Minnesota English Composition Scales: Their Derivation and Validity"; in *Educational Administration and Supervision*, vol. 7, pp. 481-99. (December, 1921.)
- Van Wagenen, M. J. "The Van Wagenen Reading Scales in History, General Science, and English Literature"; in *Journal of Educational Research*, vol. 3, pp. 314-16. (April, 1921.)

- Van Wagenen, M. J. "The Accuracy with which English Themes may be Graded with the Use of the English Composition Scales"; in *School and Society*, vol. 11, pp. 441-50. (April 10, 1920.)
- Willing, M. H. "The Measurement of Written Composition in Grades Four to Eight"; in *The English Journal*, vol. 7, pp. 193-202. (March, 1918.)
- Wilson, G. M. "Language Error Tests"; in *Journal of Educational Psychology*, vol. 13, pp. 341-49, 430-37. (September, October, 1922.)
- Witham, Ernest C. "The Use of a Composition Scale"; in *Journal of Educational Psychology*, vol. 10, p. 461. (November, 1919.)

CHAPTER VII

GEOGRAPHY AND HISTORY¹

Standardized achievement tests in geography and history less important. Although a number of achievement tests have been constructed in the field of geography and United States history, they are relatively less important than those we have considered in the preceding chapters. There are three reasons for this: (1) There is less agreement concerning the minimum essentials of these subjects. (2) For many of the outcomes of instruction intended to be engendered by the teaching of these subjects, the problem of measurement is difficult. (3) The study of most of the topics of these subjects extends over a relatively short period of time. For this reason there is much less opportunity for using diagnoses made by means of standardized tests. These points are discussed more fully in the next chapter in connection with tests for use in high schools.

When the content of a test is in agreement with the educational objectives of the school, it will be helpful in making a general survey for supervisory purposes. When using one of the tests described on the following pages for survey purposes, it is important to keep in mind that the measurement of achievements is limited, for the most part, to information. Thus we are unable to secure complete measurement of the outcomes which we endeavor to engender. This limitation is even more significant in the case of individual

¹ The account of standardized tests in the field of history is confined to those for United States history. A few tests have been devised for other divisions of history, but they can scarcely be said to have passed beyond the experimental stage.

pupils. In addition to having a value for survey purposes, the tests described should suggest to teachers types of exercises which will enable them to make their examinations more objective and more representative of their subjects.

Emphasis upon use of standardized tests may be harmful. It has been found that when a given standardized test is used, the particular outcomes which it measures tend to be given special attention by both teachers and pupils. As a result other objectives in the same subject-matter field are likely to be neglected. Information represents one important objective in both geography and history, but instruction in these subjects should engender other types of outcomes. If the use of standardized tests for measuring information in geography or history is emphasized so that these other objectives are minimized or neglected, the total effect will be harmful. Hence, when using a standardized information test in geography, history, or other content subjects, one should make certain that other objectives receive appropriate emphasis.

I. GEOGRAPHY

The problem of measurement in geography. The problem of measurement in geography is in part one of measuring the pupil's store of geographical information. It also includes the measurement of the ability of the pupil to use the facts which he knows. This use involves comparison, evaluation and organization of the facts which have been recalled. Many of the questions which are asked in the classroom are not suitable for testing purpose. It is necessary to restrict a geography test to exercises which require little writing and which can be conveniently and objectively scored. Hence it has been necessary for authors of geography tests to exercise much ingenuity in formulating suitable exercises, especially in the case of those designed to measure a pupil's abil-

ity to use geographical information. Another difficulty is encountered in determining what facts are sufficiently important to be included in the test. This phase of the problem of measurement is similar to that encountered in spelling in determining the most commonly used words. It is, however, more difficult to solve. When one attempts to measure achievement other than information, additional difficulties are encountered.

1. **The Hahn-Lackey Geography Scale.** This is a list of carefully selected questions in geography which have been classified and arranged in the form of the Ayres Spelling Scale. The questions which are approximately equal in difficulty have been printed in the same column. At the top of the column we find the per cent of correct answers given by pupils in the different grades. The following is a description of construction and selection of the exercises for this scale:

Since texts will be used by a large majority of teachers for years to come, our primary purpose was to construct a scale for the testing of the teaching of geography from textbooks. But when we realized that not one but a number of texts are being taught, we had to modify our plan. Our first modification consisted of limiting our questions to the phases of geography treated in common by six modern texts. Then we found that some of these phases were treated more fully by some authors than they were by others. A second modification of our plan was, therefore, necessary; namely, to select the common subject-matter, or, in other words, the essentials of subject-matter in each phase. In the selection of the essentials of subject-matter, the common subject-matter in these texts was largely our guide, but we also checked our exercises by principles and minimum essentials as they have been worked out by makers of geography curricula. (See 1914 and 1916 *Yearbooks of the National Society for the Study of Education*.) Over six hundred questions and exercises were selected by three teachers, covering this common subject-matter. These exercises were then examined by the authors of the scale, first, with reference to repetitions, and duplications were eliminated. They were next examined for language difficulty. The wording of many of the exercises was

changed, some of them were actually tried out on children, and in many instances technical expressions which would convey exact meaning to mature students of geography were eliminated and the ordinary language of children substituted. This is particularly true of the exercises in the lower reaches of the scale. The exercises intended for the upper reaches of the scale were not freed from technical expressions the meaning of which pupils are expected to know as evidence of geography ability. Thus we find such expressions in the scale as "the Fall Line," "climate," "continent," "natural wonders," "natural geographic barriers," "agencies," "cyclonic storms," and many others equally as technical. The exercises were examined, in the third place, as to their scope, as suggested before. Nothing was included beyond the essentials of geography. Finally, the list of exercises was revised so that it contained about an equal number of memory and thought questions.¹

This geography scale is a classified list of questions from which the teacher can select questions for a test. Since the questions in any column are equally difficult, it is best to take the questions for a test from one column. These may be given in the usual way by writing them on the board. It is better if each pupil is provided with a mimeographed copy with space left for writing in the answers. The teacher should not explain the meaning of any words used in the questions because the results will then not be comparable with the standards. Ten questions will make a test of convenient length but a larger number will make a more reliable test.

2. Courtis Supervisory Tests in Geography. One test relates to "states and important cities in the United States," and the other to the "world — oceans, continents and countries." The following description is for the former but the one on the "world" is similar in structure. There are two forms of each test.

¹ Lackey, E. E., "A Scale for Measuring the Ability of Children in Geography"; in *Journal of Educational Psychology*, vol. 9, pp. 443-51. (October, 1918.)

Form A

INSTRUCTIONS



After each State in the list write the number printed in that State on the map. Look at the first question. About what State does it ask? (Michigan.) Now look at Michigan on the map. What number is printed on Michigan? (1.) Write 1 in the space after the first question. About what State is the second question? (Ohio.) Look at the map. What number do you find on Ohio? (47.) Write 47 after the second question. What number should be written after the third question? In the same way

answer all the questions.

After each city write the number of the State in which that city is located. About what city is the first question? (Detroit.) Look at the map and find Detroit. What is the number printed in the State in which Detroit is located? (1.) Write 1 in the space after the first question. About what city is the second question? Look on the map and find what number should be written after Chicago. Write it. Write the answers to the other questions in the same way.

STATE

CITY

Questions	Number		Questions	Number	
1. Michigan ?.....			1. Detroit ?.....		
2. Ohio ?.....			2. Chicago ?.....		
3. Indiana ?.....			3. Cleveland ?.....		
4. Illinois ?.....			4. Indianapolis ? ..		
5. Wisconsin ?.....			5. Madison ?		

FIG. 17. PRELIMINARY TEST OF THE COURTIS SUPERVISORY TEST IN GEOGRAPHY. STATES AND IMPORTANT CITIES IN THE UNITED STATES

The plan of the test is to provide each pupil with an outline map of the United States showing the boundaries of the several States. Each State is given a number. The first part of the test consists of answering for each State the question, "On the map above what is the number of?" In the second part of the test the pupil is asked to give the numbers of the State in which certain cities are lo-

cated. The preliminary test which is given so that the pupil may understand just what he is to do, is reproduced in Fig. 17 to illustrate this type of test.

These tests are simple to give and require only five minutes of the pupil's time. However, it is obvious that they do not cover the whole of geography. They measure a pupil's acquaintance with certain specific geographical facts. If these facts have not received undue emphasis in the instruction which the pupils have received, scores made on these tests may be considered rough indices of their knowledge of all geographical facts which have received similar emphasis. On the other hand, if the geographical facts embodied in these tests have received any special emphasis, the measures secured will not constitute a valid index of the total achievements in this field. Hence the value of these two tests depends upon the relative emphasis which has been placed upon the informational topics of geography. If either of these tests were used frequently, the items of information which they include would probably come to be emphasized so that the scores obtained would not constitute a fair measure of the total results of the teaching of geography. If the teachers are directed to use the norms of the test as objectives as is recommended in the case of tests in the operations of arithmetic, the Courtis Supervisory Tests in Geography will have little value as measuring instruments.

3. Gregory-Spencer Geography Tests. This is a series of eight tests printed as an eight-paged booklet. A total of 111 geographical items are included. Test 1 deals with trade routes and the products which are carried over them. The exercises are like the following:

Tea	}	is shipped from Duluth to Buffalo.
Iron		
Rubber		

The pupil is asked to check the word which makes the truest

sentence. Test 2 is a group of miscellaneous exercises similar to the one above in structure. Test 3 is devoted to "causal geography" of the United States. Test 4 is similar in structure but pertains to the world. This is a sample of the exercises in Test 3:

Birmingham is one of the larger manufacturing cities of the South, because:

- it is in the cotton belt.
- it is located near limestone, coal and iron beds.
- it has an abundance of water power.

The pupil is asked to check the best reason. Test 5 deals with the location of twenty-four cities of the world. On a map of the world seventy-four cities are located and numbered. The pupil is asked to write opposite the name of each city the number which gives its location. Test 6 gives a list of descriptive phrases which apply to the twenty-four cities named in Test 5. The pupil is asked to "match" the name of each city with the descriptive phrase which applies to it. Test 7 and Test 8 deal with the countries of the world and are similar in structure to Test 5 and Test 6 respectively.

In administering the tests, the pupils are to be given all the time they need but since very little writing is required it should not require a very long period for all eight tests. The scoring is highly objective and stencils have been constructed so that it may be done very quickly. There are three forms. Unfortunately the maps are not very clear and may prove confusing to some pupils.

This series of tests is much more comprehensive than the Curtis Supervisory Tests in Geography and for this reason the limitation discussed on page 277 applies with much less force to the Gregory-Spencer Geography Tests. However, 111 items does not constitute the whole of geographical information. They are only a sample, and the validity of the

measures yielded by these tests depends upon the representative character of their content. From a partial account of the construction of the tests it appears that the authors exercised considerable care in the selection of the items included. They are intended to be representative of the "subject-matter actually being taught." The course of study in geography varies widely in different cities and in different states. Hence the content of the tests is probably not representative of the geography taught in many places. Whenever the exercises of these tests are not representative of the instruction which a given group of pupils have received, the tests will not yield valid measures of the total geographical achievements of such pupils. Since the same tests are to be used in Grades VI, VII, and VIII it is doubtful if they are equally representative of the instruction in these three grades.

This series of tests is intended to have a diagnostic function. Separate measures are secured for each of the tests. It is expected that a study of these measures will enable a teacher or a supervisor to determine if the proper emphasis is being given to the different phases of geography which these tests measure.

4. Posey-Van Wagenen Geography Scales. Information Scale R (General) consists of a series of informational questions arranged in ascending order of difficulty. Division I is for Grades V and VI and Division II for Grades VII and VIII. There is a corresponding Information Scale S and separate scales for the continents: A, United States and North America; F, Europe; K, South America, Asia, and Africa. The questions are of such a nature that very little writing is necessary in answering them. The scoring is rather highly objective though somewhat less so than for the Gregory-Spencer Geography tests. The following are samples of the questions from Information Scale R, Division I:

7. Name the two largest cities of Europe.
20. Name three island possessions of the United States.
29. After the names of each of these animals write the name of the continent in which it is found.
 1. Giraffe
 2. Kangaroo
 3. Llama
 4. Zebra

Thought Scale S consists of a different type of exercises arranged in ascending order of difficulty. There are two divisions as in the Information Scale R. The following are samples of the exercises:

4. In South Carolina and Georgia only the coarser kinds of cotton cloth are manufactured in large quantities, while in Massachusetts the finer grades are manufactured. In which State would you expect the factory hands to be paid the highest wages?

19. The principal manufactures of St. Louis are tobacco, flour, and meat products. By what sort of an industrial region is it surrounded?

27. Other conditions being equal water is lost through the leaves of plants and trees more rapidly during warm than during cold weather, more rapidly during windy than during calm weather, more rapidly from large leaf surface than from small leaf surface and more rapidly from trees that reach up into the drier layers of air than from low ones. Southern Europe is much warmer and has less rainfall than central Europe.

(a) In which section would you expect to find the heaviest forests?

.....
(b) Would you expect the olive trees of southern Europe to be low or tall?

.....
(c) Would you expect them to have small leaves or large ones?

.....

No account of the construction of these tests is available, but since they are in the form of difficulty scales it is reason-

ably certain that the selection of the exercises was partly on a statistical basis; that is, from a preliminary list those exercises were chosen which were found to possess appropriate degrees of difficulty. This procedure generally results in the content of a test being less representative of the subject-matter field than when the exercises are selected solely on the basis of their importance when judged by our major educational objectives. Incidentally it may be noted that questions relating to items which have received little or no attention in the instruction will tend to be difficult because many of the pupils have not had an opportunity to learn them.

The comments upon the limitations of the Gregory-Spencer Geography Tests apply to these scales also. It is likely that they apply to greater extent because the Posey-Van Wagenen Geography Scales appear to be less representative of instruction in geography. In addition the latter scales are more difficult to use. The calculation of a pupil's score will not be understood by many persons and hence will be found difficult.

The Thought Scale S is an attempt to measure the ability of children to think about geographical questions. For some pupils it will doubtless be a thought test, but for one who remembers the answer to the question, there will be no thinking. For such a pupil the exercise is merely one calling for a specific fact. Although this constitutes a limitation of the scale as an instrument for measuring ability to think, it is inherent in the working of the human mind rather than in the scale. Whenever we have a ready-made answer for a question we do not think. It is a fact question. For another person who does not have a ready-made answer, the question is one calling for thought.

5. Witham Standard Geography Tests. There are eight tests in this series: (1) The World, (2) United States, (3)

South America, (4) Europe, (5) Asia, (6) Africa, (7) North America, and (8) Commercial Geography. The first seven are devoted to informational questions relating to countries, cities, seas (oceans), rivers, mountains, industries, and products. Many of the questions relate to maps which are not good. The tests are announced as having a diagnostic function. A separate score is given for each of the major divisions of each test and a graph sheet is provided which assists in interpreting the scores of a class.

II. UNITED STATES HISTORY

The problem of measurement in history. It is difficult to define performances which may be accepted as evidence of many of the desired achievements in the field of history. Certain historical facts are to be memorized. Evidence of the extent of such achievements may be secured by asking questions which call for the specific facts which students are expected to memorize. For example, knowledge of the name of the first President of the United States may be measured by asking the question, "Who was the first President of the United States?" This question calls for a definite fact. Only one answer is correct. All others are wrong. If the problem of measurement in history involved the construction of tests to measure only achievements of this type — that is, memorized facts — the only difficulty would be the determination of a minimum essential list of such facts. Probably no teacher of history would agree that the memorization of facts represents the total achievements for which we strive in history. We expect pupils to acquire a mass of general information which they may not remember in detail and are unable to express in concise language, but which functions in such situations as the interpretation of current events mentioned in their general reading. Such achievements may be termed attitudes and perspectives. In addition,

many ideals are to be engendered in the teaching of history, particularly that of the United States.

It is difficult to devise a type of exercise which will secure a performance that may be used as evidence of these more subtle achievements — attitudes, perspectives and ideals. What must a pupil do or say in order to demonstrate that he has acquired “the ideal of patriotism,” “the historical attitude,” or “the historical perspective for understanding an account of the causes of the Civil War”? For certain achievements it is very doubtful if a satisfactory performance could be secured upon demand from all pupils. So far no one has been successful in devising a test which would measure these more subtle achievements.

In addition to tests for measuring historical information, a few tests have been devised to measure ability to use historical information in answering thought questions. As we have already pointed out, whether a given exercise constitutes a “thought question” depends upon the person who is answering it as well as its wording. What constitutes a thought question for one may be only an information question for another. It all depends upon whether he has a ready-made answer. If he does it is an information question. Thus, a “thought test” is likely to be an information for some pupils or even for entire groups of pupils.

It has been stated that no satisfactory history tests have yet been devised. In one sense this is true, but it is believed that a consideration of the available tests will be helpful. Many of the authors have exhibited ingenuity in devising exercises which permit of objective scoring and which require a minimum time to answer. These exercises should be suggestive to teachers of history in the formulation of their own tests and examinations.

1. Hahn Scale for Measuring Ability of Children in History. This scale is very similar in structure to the Hahn-

Lackey Geography Scale described on page 274. It is limited to the history of the United States and is designed for use in the seventh and eighth grades. It is to be used as a source of questions by the teacher.

2. **Harlan Test of Information in American History.** The items of information called for by this test are found in practically all American history textbooks, being based on the study by Bagley and Rugg of twenty-three textbooks to determine the content of American history.¹ The credit to be given for answering each question has been determined and a score card is furnished with the test so that the marking of test papers may be uniform. The test consists of ten exercises. The first, fourth, and ninth are reproduced to illustrate their nature.

EXERCISE I

At the right of the page are the names of some men mentioned in American history. Fill in blanks with the names which properly belong there.

Score.....1. America was discovered by.....
near the close of the fifteenth century.

- | | |
|--|--|
| 2. The name of the man who is supposed to have discovered the Pacific Ocean is..... | Jefferson
Cornwallis
William Penn |
| 3. The first President of the United States was..... | Lafayette
Patrick Henry
Columbus |
| 4. is the name of a distinguished Frenchman who aided the colonists in securing their independence. | Benj. Franklin
Washington
John Cabot
Balboa |
| 5. surrendered to the colonial troops at Yorktown. | |

¹ Bagley, W. C., and Rugg, H. O., *The Content of American History as Taught in the Seventh and Eighth Grades.* (University of Illinois, School of Education, Bulletin no. 16, 1916.)

EXERCISE IV

Tell the very first thing you would do under each of the following conditions; also what you would do next:

- Score.....1. If a neighbor were to present to you for your signature a petition to have some man removed from public office —
 What would you do first?.....
 Would you sign the petition?.....
2. If a man imprisoned in the county jail for some serious crime should be taken out by a mob with the intention of hanging him —
 What ought to be done first?.....
 Then what?.....

EXERCISE IX

The following topics represent matters of importance in the history of the United States. State definitely of what significance each has been.

- Score.....1. Articles of Confederation.....

 2. Mason and Dixon's line.....

 3. Monroe Doctrine.....

 4. The Tariff.....

The total amount of credit allowed for the ten exercises is 100. Over two thousand pupils in the seventh and eighth grades have been tested at the end of the respective years. The median scores are: Seventh grade, 48; eighth grade, 67. These standard scores are the scores which should be made by the average or median (middle) pupils. Consequently, when translating these scores into school marks, a score which is standard should be given the "school grade" which represents the average pupil.

The different exercises call for different types of information: dates, names, causes, meanings of historical terms, etc. By tabulating separately the credits earned on the different exercises a teacher may learn which phases of history need additional emphasis.

3. Van Wagenen American History Scales.¹ Van Wagenen has constructed three types of tests: Information Scale A and Scale B, Thought Scale A and Scale B, and Character Judgment Scale A, Scale B and Scale L. In these tests the exercises are arranged in ascending order of difficulty. Their intended function is implied in their titles. In Van Wagenen's Information Scale A there are a number of questions of the usual type and some exercises which are different. Three of these are reproduced below:

10. Arrange these events in the order in which they occurred by putting a "1" before the event that occurred first, a "2" before the event that occurred second, and so on until you have put a "5" before the event that occurred last.

Struggle between the French and English for control in America.

Rise and growth of the United States as a nation.

Discovery of America.

Settlement of America by European nations.

Struggle of the American colonies against European control.

28. Which of these men won each of the following battles:

Dewey, Perry, Grant, Farragut, Morgan, Taylor, Thomas:

Battle of Cowpens?.....

Battle of Mobile?.....

Battle of Manila?.....

Battle of Buena Vista?.....

Battle of Nashville?.....

Battle of Vicksburg?.....

Battle of Lake Erie?.....

30. Put a check mark in front of each of the following things which the Southern States were in favor of between 1840 and 1850.

¹ As the galley proof is being read, a letter from Professor Van Wagenen states that these scales have been revised. The form has been changed, but many of the exercises are the same.

.....Wilmot Proviso.	The "gag rule" or suppression of abolition petitions in Congress.
.....William Lloyd Garrison's "The Liberator."	Admission of California as a state.
.....Protection of slavery in the territories.	Annexation of Texas.
.....Free Soil Party.	Protective tariff on manufactured goods.

The following exercises are taken from Van Wageningen's Thought Scale A:

6. In 1800, Spain gave Louisiana up to France. The United States, fearing that France might set up a colony and control the Mississippi River, was anxious to get Louisiana. In 1803, Napoleon of France feared that Great Britain was about to seize his American territory.

What would you expect Napoleon to do?

15. In 1649, Oliver Cromwell became the ruler of England, the King Charles I, having been driven from the throne and put to death. The Royalists, who had favored the king, belonged to the Church of England. During the next few years a large number of people left England to settle in America.

(a) Who do you think these new settlers were?

(b) To what colony in America would these people be most likely to go?

20. In the rural communities in 1850 the children had an opportunity to learn many things in the home which they could not learn in the city homes. When a tax was raised upon property the people in the rural communities, who owned their farms for the most part, had to pay a larger proportion of the tax than the workingmen of the cities.

(a) When in 1849 and in 1850 the bill for free public schools was submitted to a vote of the people of New York State, what way would you expect the workingmen of the cities to vote?

Why?

(b) What way would you expect the farmers in the rural communities to vote?

Why?

In his Character Judgment Scale, Van Wagenen attempts to have the pupil infer concerning a man's character. A pupil is to give his response by marking a certain descriptive word. The second exercise from the Character Judgment Scale A is quoted.

2. After the British troops were driven from Boston by Washington's clever maneuvers, New York became the scene of war. Here the military situation was most serious. The British numbered 25,000 well-equipped troops, with a large number of cannon, generous stores of ammunition and even ships at their command. The Americans numbered but 14,000 poorly-equipped and ill-fed men. Washington saw that he must have certain news of the enemy; he must know exactly the number of their troops and how they were posted in the defense of New York. He needed a spy, — one who would enter the lines of the British, learn all he could, and return with the information to the commander-in-chief. Then Washington would know the place and time to make an attack.

With the alert eyes and ears of hundreds of enemies about him, the spy rarely escapes detection. If discovered, he is not shot, but hanged. When Washington asked for volunteers, Nathan Hale consented to enter the British lines as a schoolmaster who was disgusted with the American cause.

Draw a line under the three of the following words which you think best describe the action of Nathan Hale.

cowardly	prudent	ignoble	fearless	daring
treacherous	cautious	courageous	selfish	faithless

Criticism of these scales. Van Wagenen states that "Information Scales A and B are designed to measure the range of information from the standpoint of quantity and difficulty of comprehension." Thought Scales and Character Judgment Scales were designed as companion scales. Hence, Van Wagenen attempted to distinguish between information and other outcomes of the teaching of American history. He even goes so far as to say that the entire list of scales is not intended to cover the whole field of mental activity involved in the study of history.

One feature of these tests is the inclusion of a provision for allowing partial credit for exercises partly right. Although Van Wagenen has prepared detailed directions for scoring, it is sometimes difficult to determine the credit which a pupil should receive. The author has given the following reliability coefficients: Information $.71 \pm .01$, Thought $.74 \pm .01$ and Character Judgment $.83 \pm .01$.

We have already commented upon one difficulty in constructing thought questions. It is illustrated in the Van Wagenen Thought Scales. For example, consider the exercises quoted from Van Wagenen's Thought Scale A. The answer required in each case is a historical fact. In Exercise No. 6 (see p. 287) a pupil might "expect" Napoleon to defend Louisiana or to sell it to Great Britain. The exercise does not give sufficient information about the countries concerned in order to make possible a rational judgment concerning Napoleon's actions. Since the only answer accepted as correct is that Napoleon did sell Louisiana to the United States, the question becomes essentially a fact question.

The inadequacy of the information furnished by the exercise is even more forcibly illustrated by the fifteenth exercise which is also quoted on page 287. A pupil cannot determine to what colony the people who left England after 1649 would be most likely to go unless he possesses information concerning the American colonies. Again the answer required is a historical fact.

In Van Wagenen's Character Judgment Scale A, a pupil might make deductions concerning the character of Nathan Hale from the description given, but it is likely that most pupils would be able to answer this question from memory. In other words, for most pupils this question becomes one of information rather than character judgment.

Similar criticisms can be made with reference to practically all the exercises in Van Wagenen's scales. Some of

them doubtless require a pupil to exercise abilities other than mere information, but for a surprisingly large number the answer required is a historical fact. Even if the exercise presents all of the information which the pupil needs to make an inference, pupils are not likely to go through with the higher mental processes if they are able to give the answer required from memory. To eliminate the possibility of an exercise becoming a mere information exercise it would be necessary to have the tests deal either with topics of history not studied, and for which they possessed no information, or with fictitious items.

4. Barr Diagnostic Tests in American History. This is a group of five tests designed to measure separately (1) comprehension of material read, (2) chronological judgment, (3) judgment of evidence, (4) evaluation of importance of facts, and (5) understanding of cause and effect relationships. The exercises of Test 1 are similar to those found in certain of our silent reading tests. Samples of the exercises in the other tests follow.

TEST II. CHRONOLOGICAL JUDGMENT

1. Indicate the chronological order (time order) in which the following men lived by placing a figure one (1) before the individual who lived first, a figure two (2) before the individual who lived second, and so on through the list.

- (a) Patrick Henry
- (b) Robert E. Lee
- (c) John Cabot
- (d) James Monroe
- (e) Roger Williams

3. Put a cross (X) before three of the following events which occurred about the same time (same general period). It is not necessary that these events should have taken place at exactly the same date.

- (a) The purchase of Florida
- (b) The Emancipation Proclamation

- (c) The settlement of Plymouth
- (d) The Declaration of Independence
- (e) The building of the Panama Canal
- (f) The battle of Vicksburg
- (g) The building of the Erie Canal
- (h) The Spanish-American War
- (i) The reelection of Lincoln
- (j) The discovery of America

4. From the list at the right of the page select the names of three men who were prominent in each of the following periods. Write their names under the name of the period in which they were prominent.

(a) Discovery and exploration

.....

- (1) Abraham Lincoln
- (2) Benjamin Franklin
- (3) Henry Clay

(b) Revolutionary War period

.....

- (4) Patrick Henry
- (5) Christopher Columbus
- (6) Champlain
- (7) Thomas Jefferson

(c) The struggle over slavery
 (1840-1866)

.....

- (8) John C. Calhoun

9. Indicate approximately how long ago the following events occurred by placing one of the following terms before each: decade, two decades, century, one half century, one quarter century, two centuries, etc. Do not answer by giving the date.

- (a) Monroe Doctrine, Declaration of the
- (b) The annexation of Texas
- (c) Bacon's Rebellion
- (d) The completion of the Panama Canal
- (e) The battle of Lexington

TEST III. JUDGMENT OF EVIDENCE

5. (A) "On Feb. 24, 1779 Lieut. Governor Hamilton, who was besieged by Colonel Clark, proposed a three day truce. Colonel Clark refused the truce but expressed his willingness to accept the

surrender of the garrison and suggested a conference. The meeting was held in one of the churches of Vincennes."

(B) "Colonel Clark's compliments to Lieutenant Governor Hamilton and begs to inform him that he will not agree to any terms other than Mr. Hamilton's surrendering himself and garrison prisoners at discretion. If Mr. Hamilton is desirous of a conference with Colonel Clark, he will meet him at the church with Captain Halm, Feb. 24 1779." Signed. G. R. Clark.

(a) If you were writing a historical account of the siege described above, upon which of these two statements, everything else being equal, would you base your account?

.....
(b) Give one reason in support of your conclusion.
.....

TEST IV. EVALUATION OF IMPORTANCE OF FACTS

1. Put a cross (X) before the event in the following list which has been of the greatest importance in American History.

- (a) The founding of Philadelphia
- (b) The purchase of Alaska
- (c) Jay's treaty with England
- (d) The Declaration of Independence
- (e) The election of Andrew Jackson

8. Put a cross (X) before the fact in the following list which has been of the greatest importance in the industrial development of the United States.

- (a) The Kansas-Nebraska Act
- (b) The formation of the Standard Oil Company
- (c) The introduction of large-scale production into manufacturing
- (d) The financial crisis (panic) of 1907
- (e) The introduction of Bessemer steel

TEST V. UNDERSTANDING OF CAUSE AND EFFECT RELATIONSHIPS

2. In the left-hand (first) column below is a list of causes. At the right-hand (second) column is a list of results. Take your pencil and draw a line connecting each cause with the proper result.

- | | |
|---|---------------------------------------|
| (a) The sinking of the Maine | (a) Territorial claims on Oregon |
| (b) The growth of industrial combinations | (b) The Mexican War |
| (c) The annexation of Texas | (c) The growth of abolition sentiment |
| (d) The Dred Scott decision | (d) The Spanish-American War |
| (e) The Lewis and Clark expedition | (e) The Sherman Anti-Trust Law |

5. Put a cross (X) before each of the following facts which was either directly or indirectly a cause of the REVOLUTIONARY WAR.

- (a) The Articles of Confederation
- (b) The Boston Tea Party
- (c) The Lewis and Clark expedition
- (d) The Stamp Act
- (e) The Intolerable Acts
- (f) The invention of the steamboat
- (g) The election of Washington
- (h) The Boston Massacre

These five tests make a rather elaborate series and hence require considerable time to give and score. The scoring is not completely objective and involves some complications.

Gregory Tests in American History. Part 1 is a miscellaneous list of forty questions calling for specific facts and dates. Each of the other four parts is devoted to a period in our history. Part 2, the period of National growth from 1789 to 1829; Part 3, the period of sectional disputes and Civil War; Part 4, the period of reconstruction and National development; Part 5, the period from 1900 to 1922. In the exercises of these four parts the pupil is given three statements from which he is asked to select the one which makes the truest sentence. The four parts include a total of sixty exercises. The following is a sample exercise.

THE KANSAS-NEBRASKA BILL

.....reestablished the Missouri compromise of 1820 which made all territory west of Missouri and north of thirty-six degrees and thirty minutes free.

.....extended the principle of "Squatter Sovereignty" to the territory of Kansas and Nebraska.

.....provided that Kansas should enter the Union as a slave State and Nebraska as a free State so as to maintain the balance of power.

A feature of the test is the small amount of writing required of the pupil and the objectivity of the scoring. There is no information available concerning the method of selection of the items of information included in the test.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. Why are standardized achievement tests less important in geography and history than in the tool subjects, arithmetic, reading, spelling and language?
2. Which of the tests described in this chapter will be most useful to the teacher?
3. What is a thought question? What is a memory question?
4. How does the method which Hahn and Lackey used, in determining the minimum essentials in geography, compare with the method used by Ayres in Spelling? Which method will give the most dependable results? Why?
5. In what respects does the problem of measurement in geography and history differ from the problems of measurement discussed in the preceding chapters?
6. Is it likely that sometime we will have standardized tests in geography and history which are as satisfactory as those we now have in arithmetic and silent reading? Why?

SELECTED BIBLIOGRAPHY

- Barthelmess, Harriet M. "Geography Testing in Boston"; in *Journal of Educational Research*, vol. 2, pp. 701-12. (November, 1920.)
- Buckingham, B. R. "A Proposed Index of Efficiency in Teaching United States History"; in *Journal of Educational Research*, vol. 1, pp. 161-71. (March, 1920.)

- Byrne, Lee. "Using Home-Made Tests in High Schools"; in *The School Review*, vol. 80, pp. 536-46. (September, 1922.)
- Clark, Marion G. "A Study in Testing Historical Sense in Fourth and Fifth Grade Pupils"; in *The Historical Outlook*, vol. 14, pp. 147-50. (April, 1923.)
- Courtis, S. A. "Measuring the Effects of Supervision in Geography"; in *School and Society*, vol. 10, pp. 61-70. (July 19, 1919.)
- Cram, Fred G. "The Scores of a Group of Rural Teachers on Courtis Standard Supervisory Test in Geography"; in *Journal of Educational Research*, vol. 2, pp. 515-17. (June, 1920.)
- Gibson, O. H. "Existing Standard Tests in History"; in *The Historical Outlook*, vol. 12, pp. 324-26. (December, 1921.)
- Gregory, C. A., and Spencer, Peter L. "A Geography Test for the Sixth, Seventh, and Eighth Grades"; in *School and Society*, vol. 15, pp. 452-56. (April 22, 1922.)
- Griffith, G. L. "Harlan's American History Test in the New Trier Township Schools"; in *School Review*, vol. 28, p. 697. (November, 1920.)
- Harlan, Chas. L. "Educational Measurements in the Field of History"; in *Journal of Educational Research*, vol. 2, pp. 849-53. (December, 1920.)
- Lackey, E. E. "A Scale for Measuring the Ability of Children in Geography"; in *Journal of Educational Psychology*, vol. 9, pp. 443-51. (October, 1918.)
- Mathewson, Chester A. "The Use of the Hahn-Lackey Geography Scale"; in *Journal of Educational Psychology*, vol. 9, pp. 467, 581-87. (October, November, 1918.)
- Myers, A. F. "Harlan's Test of Information in American History. Re-classification of Children on Basis of Tests in Port Clinton Schools"; in *Journal of Educational Method*, vol. 1, pp. 24-25. (September, 1921.)
- Odell, C. W. "The Barr Diagnostic Tests in American History"; in *School and Society*, vol. 16, pp. 501-03. (October 28, 1922.)
- Parker, Edith P. "A Few Suggestions for Informal Testing in Geography"; in *Elementary School Journal*, vol. 23, pp. 444-47. (February, 1923.)
- Pratt, Orville C. "Spokane United States History Test"; in *Journal of Educational Research*, vol. 3, pp. 155-57. (February, 1921.)
- Pressey, S. L. "Scale of Attainment No. 2 — An Examination for Measurement in History, Arithmetic, and English in the Eighth Grade"; in *Journal of Educational Research*, vol. 3, pp. 359-69. (May, 1921.)
- Rugg, E. U. "Character and Value of Standardized Tests in History"; in *School Review*, vol. 27, pp. 757-71. (December, 1919.)
- Van Wagenen, M. J. *Historical Information and Judgment in Pupils of Elementary Schools*. Contributions to Education, no. 101. (New York: Teachers College, Columbia University, 1919.)

- Van Wagenen, M. J. "The Van Wagenen Reading Scales in History, General Science, and English Literature"; in *Journal of Educational Research*, vol. 3, pp. 314-16. (April, 1921.)
- Witham, Ernest C. "Standard Geography Test — The World, for Fifth Grades"; in *Journal of Educational Psychology*, vol. 9, pp. 432-42. (October, 1918.)
- Witham, Ernest C. "A Test on Roman History"; in *A School Review*, vol. 30, pp. 489-90. (September, 1922.)

CHAPTER VIII

HIGH-SCHOOL TESTS

High-School Tests described in preceding chapters. Several of the tests described in the preceding chapters are designed to be used in the high school as well as in the elementary school. This is true of certain tests in silent reading, English, and United States history. The important tests in these subjects have been described in the preceding chapters. Some tests designed for use in the elementary school have been used above the eighth grade. However, when this is done, one should recognize that the test is being given to pupils for whom it was not designed. In most cases the test will be only partially satisfactory.

Standardized tests described in this chapter. In this chapter we shall describe certain standardized tests in mathematics, Latin, modern languages, and science. No attempt is made to describe all of the tests which have been devised in the fields of these subjects. This is particularly true in the case of science. The principal reason for limiting the number of tests treated is that given in the introduction to the treatment of standardized tests for geography and history. Standardized tests in other school subjects are not included, although several have been devised. In the bibliography at the end of this chapter the more important references for such tests are given under the heading "Other Subjects."

Limitations of achievement tests for use in the high school. In addition to the general limitations of educational tests, certain ones are introduced by the nature of the educational objectives of the high school. The function of an achieve-

ment test is to yield measures of the extent to which pupils have attained certain objectives which have been set for them. For example, if we accept as an objective in the field of spelling the ability to spell correctly a certain list of one thousand words, an achievement test for this field should yield a measure of the ability of pupils to spell these one thousand words correctly. Similarly in the field of arithmetic an achievement test implies certain objectives and its function is to measure the degree to which pupils have achieved these objectives.

In the elementary school we have reached a fairly definite agreement upon certain minimum essentials in such subjects as arithmetic, silent reading, spelling, and handwriting. In the high school there is far less agreement in regard to the objectives. In history, for example, authorities differ in regard to the details of minimum essentials except in the case of a few of the most formal items. In fact, it does not appear to be essential that the content of such subjects as history should be fixed to the extent that the content of handwriting, the operations of arithmetic or spelling, should be. It seems likely that two teachers of European history might be equally efficient in realizing the ultimate educational objectives, but vary widely in the emphasis which they place upon different topics. Indeed, it is conceivable that they might exhibit considerable lack of agreement with respect to the topics included in the course.

When certain exercises are chosen for a test which is to be printed and offered for universal use, it is implied that these exercises should rightfully be included in the educational objectives of that subject. Hence, agreement upon the group of educational objectives to be attained in the field of a subject is a prerequisite for the construction of a satisfactory achievement test in that field. Because of the lack of agreement in regard to the details of educational objectives in

high-school subjects very definite limitations are placed upon the achievement tests.

Another limitation is placed upon the measurement of achievement in the high school by the nature of the outcomes of instruction. In the elementary school skills and memorized facts are prominent among the desired outcomes. The pupil is expected to memorize many facts in arithmetic, spelling, geography, etc., and to become skillful in such activities as calculation in arithmetic, spelling, silent reading, and oral and written expression. In the high school the engendering of ideals, attitudes, and perspectives becomes prominent. These outcomes of instruction are much more subtle than skills or memorized facts. They are much more difficult to measure. It should be frankly recognized that at the present time we are not able to measure them as satisfactorily as we can skills and memorized facts.

Prognostic tests most valuable for use in high schools. It is a well-known fact that a large per cent of high-school students fail in the subjects which they undertake. Some of these students do not have the general intelligence to do the work that is required of them. Others lack the special ability required for a given subject. Some have not acquired a good technique of study. A large per cent of these failures could probably be avoided by advising students not to undertake subjects for which they are not fitted. The most valuable function of standardized tests in the high school is to yield measures which are prognostic of a student's probable success in the various subjects. Most tests are prognostic to at least a slight extent, but a number have been devised for this particular purpose.

General intelligence tests have for their function the measurement of a student's general capacity to do the work of the school. In addition to these, we have a few tests whose

function is prognostic for certain school subjects, as Rogers Tests of Mathematical Ability, Van Wageningen Reading Scales for History and General Science, and Handschin Predetermination Tests for Foreign Languages. These tests may be given at the beginning of a course or, in a few cases, before the student undertakes the work.

Ability to read silently is a prerequisite to effective study in a number of high-school subjects. In such cases a silent reading test has a general prognostic function. Students who are unable to make satisfactory scores on suitable silent reading tests will probably do unsatisfactory work in history, literature, science, and other subjects which involve a large amount of textbook study. Hence silent reading tests are useful for general prognostic purposes.

Little opportunity for diagnosis of high-school students with reference to achievement. The possibility of diagnosing students with respect to their achievements is not the same in all grades of the school. A diagnosis cannot be made until students have had some opportunity to achieve. They must have received some instruction on the topic before diagnosis is possible. In the elementary school the students pursue a number of subjects over a period of several years. For example, they study silent reading in all grades. By repeated drill they are trained to be fluent readers. Much the same situation exists in spelling, handwriting, and arithmetic. In the field of each of these subjects there is abundant opportunity for diagnosis with respect to achievement before the period of learning is completed.

In the high school, however, the situation is materially different. As a rule, when a topic has been studied, a student does not return to it except incidentally or in the course of review. There are certain exceptions such as the operations of algebra, and reading of a foreign language in which the engendering of skills extends over several months or even

a longer period. However, for the most part high-school students engage in the study of topics on which they do not receive continued training. Hence, a diagnosis with respect to achievement is, in general, impossible until instruction in the subject has been practically completed. Then a diagnosis has only a limited usefulness. Until we have agreed more completely upon the particular educational objectives to be attained and have modified our plan of education so that the instruction on a topic will extend over a longer period of time, teachers must necessarily make their diagnoses in other ways than by the use of standardized tests.

Purposes to be realized in the use of educational tests in high schools. Educational tests probably render the greatest service to high-school teachers and principals in connection with the educational and vocational guidance of students. They are helpful also in the classification of students. For both of these purposes tests of general intelligence are probably of greatest value. Educational tests are much less useful for evaluating the efficiency of a high school than of an elementary school. Until we have arrived at an agreement concerning the particular objectives to be attained we cannot hope to secure from the scores yielded by even our best achievement tests more than a very rough indication of the efficiency of the high school. As indicated above, standardized tests may be used for diagnosis and other instructional purposes in certain subject-matter fields.

I. MATHEMATICS

Problem of measurement in algebra. The first difficulty which one encounters in the construction of a group of tests in elementary algebra is the determination of the types of exercises to be included. In arithmetic it is obvious that the fundamental operations are addition, subtraction, multi-

plication, and division. These are the operations which are used in solving problems. In elementary algebra the situation is not the same. Most algebra texts treat each of these operations as separate topics, but for the most part the use of them in solving problems is limited to the manipulations required for determining the values of the unknown quantities in simple, simultaneous, and quadratic equations. In addition there are many other topics such as removal of parentheses, factoring, exponents, evaluation of formulæ, and reduction of radicals. Thus one finds a large number of types of exercises in the field of elementary algebra. There are marked differences of opinion concerning the types which are sufficiently important to be included in the fundamentals of algebra.

Three methods have been used by test-makers in determining the "fundamental operations" of elementary algebra. Rugg and Clark (see page 306) chose those operations "represented by currently used textbooks." Douglass (see page 303) secured the judgment of fifty-nine members of the Mathematical Association of America and accepted as "fundamental" those operations upon which these persons most nearly agreed. In both the Monroe Standard Research Tests in Algebra and the Illinois Standardized Algebra Tests (see page 305) the writer has proceeded upon the thesis that since problems were solved by means of the equation, the "fundamental operations" of elementary algebra are those which are required in the solution of equations. According to this thesis, one set of fundamental operations would be derived from the simple equation, another from simultaneous equations, and a third from the quadratic equation. This phase of the problem of measurement is merely one of determining the minimum essentials or most important topics of elementary algebra. As we have pointed out in connection with our consideration of

other school subjects, this is a prerequisite step in test construction.

It is also necessary to consider the type of measurement which is desired. Some of the available tests are designed to measure a student's "power" to do certain types of exercises. Other tests are designed to measure a student's skill; that is, his rate and accuracy. As we have pointed out in other places these two types of tests imply different concepts of the objectives to be attained in the teaching of a school subject. In selecting an algebra test one should choose the type of test which is in agreement with his objectives.

Douglass Standard Diagnostic Tests for First-Year Algebra. Series A consists of four tests: (1) addition and subtraction, including collection of terms; (2) multiplication, ranging from $3b$ times 2, to the product of two trinomials; (3) division; (4) solution of simple equations. Series B consists of seven tests: (1) fractions; (2) factoring; (3) formulæ and fractional equations; (4) simultaneous equations; (5) graphs; (6) square roots, exponents, and radicals; (7) quadratic equations. Each test consists of ten exercises arranged in ascending order of difficulty. The time allowed for each test is sufficient for the student to attempt all of the exercises. Thus no measure of rate is secured.

Preparatory to the construction of these tests Douglass sent a questionnaire to one hundred members of the Mathematical Association of America in which they were asked to indicate those topics which they considered "the most elementary and important fundamentals" of first-year algebra. The topics included in the above tests are those which were most frequently mentioned in the fifty-nine replies which were received. With respect to the selection of the exercises for the tests of Series A the author states that he attempted "to secure a variety of exercises testing the various types

of difficulty under each fundamental [topic], thereby providing opportunity for diagnosis and at the same time providing opportunity for differentiation of different degrees of ability by the inclusion of exercises of varying degrees of difficulty." Presumably the same criterion governed the selection of the exercises for the tests of Series B. The difficulty of each exercise has been determined and a pupil's score is the sum of the difficulty values of the exercises which he does correctly.

Series A is announced as being "designed especially for testing the *fundamentals* of algebra." Series B is intended to measure ability in certain "additional processes." In the opinion of those replying to the questionnaire these are notably less important processes. As the title indicates, the tests are intended to fulfill diagnostic function; i.e., to measure separately the ability of pupils in each process and also to show the ability of pupils to deal with each of the "various types of difficulty" within the process. In order to secure the fulfillment of the latter function it is necessary to make a tabulation of the responses for each exercise. As the author points out, the brevity of the tests imposes certain limitations upon the fulfillment of their intended diagnostic function. Although two forms of these tests are available, no data have been reported concerning their reliability. This information would be useful in judging their validity.

Hotz Algebra Scales. Each of these scales consists of exercises arranged in ascending order of difficulty. The scales in Series A consist of exercises taken from Series B and equally spaced upon the scale of difficulty. This makes the increase in difficulty gradual which is not true in Series B. There are five scales in each series.

1. Addition and Subtraction.
2. Multiplication and Division.
3. Equation and formula.
4. Problems.
5. Graphs.

In selecting the exercises for his scales Hotz attempted to include "only those exercises and problems most commonly found in recent textbooks." From a preliminary list made upon this basis, those exercises were rejected which were found unsatisfactory for testing. "Some of them were not clearly stated, others ran very unevenly as seen by the fact that they were often solved more frequently by classes having little training in algebra than by more mature classes, and still others tended to group themselves too much about the same point upon a scale of relative difficulty."

The time allowed is sufficient for practically all students to try all of the exercises that they are able to do correctly. A student's score is the number of exercises solved correctly.

Illinois Standardized Algebra Tests. Each of the four tests of this group is limited to a single type of simple equation. The general character of these types of equations may be illustrated as follows:

$$\text{Test I: } \pm ax \pm bx = \pm c$$

$$\text{Test II: } \pm ax \pm c = \pm bx \pm d$$

$$\text{Test III: } \pm k(\pm ax \pm c) = \pm bx \pm d$$

$$\text{Test IV: } \pm \frac{\pm ax \pm c}{n} = \pm \frac{\pm bx \pm d}{\pm m}$$

The difficulty of an equation of a given type depends in part upon the particular combination of signs which occurs. In each test four combinations which were judged to be most typical were selected and the equations representing these combinations were arranged according to the "cycle principle." For example, in Test I the following combinations were used: $+$ $-$ $+$ $-$, $-$ $+$ $+$ $-$, $+$ $+$ $-$ $-$, and $+$ $-$ $-$ $+$. The first combination of signs is found in the first equation, the fifth, the ninth, the thirteenth, etc. Each test consists of twenty equations. The time limits are set so that a measure of the rate of work is secured. The group of

four tests can easily be administered within a forty-minute class period. A pupil's score is the number of exercises attempted and the number right.

This series of algebra tests differs from the others described in this chapter in the concept of the fundamental operations which it represents. The authors proceeded upon the thesis that the equation represents the fundamental operations of elementary algebra. This thesis is apparently not in agreement with the content of our commonly used textbooks and it does not agree with the expressions of opinion which Douglass secured. However, this evidence does not necessarily disprove the thesis. Certainly the equation is the tool which we use in solving problems. Except in doing formal exercises, a student in elementary algebra meets few if any demands for addition, subtraction, multiplication, division, and factoring except as they occur in solving equations. A similar series of tests is needed for quadratic equations and possibly one for simultaneous linear equations.

Rugg and Clark Tests in First-Year Algebra. There are sixteen of these tests.

- | | |
|----------------------|--|
| 1. Collecting terms. | 9. Clearing of fractions. |
| 2. Substitution. | 10. Fractional equations. |
| 3. Subtraction. | 11. Practical formulæ. |
| 4. Simple equations. | 12. Quadratic equations. |
| 5. Parentheses. | 13. Simultaneous equations. |
| 6. Special products. | 14. Radicals. |
| 7. Exponents. | 15. Graphs. |
| 8. Factoring. | 16. Quadratic equations
(irrational roots). |

This elaborate group of tests is the result of four successive formulations (completed in 1917) and is intended to include all of the fundamental operations "represented by currently used textbooks." In each test the important types of exercises have been "arranged rigidly in rotation." This is the

"cycle principle." It results in a pupil meeting each type of exercise an equal number of times and as a result the necessity of weighting the exercises within a cycle for differences in difficulty is eliminated. The time limits have been set so that "no student can quite finish the test in the time given, but that all can do a considerable number" of exercises. The pupil's score is the "number right" and the "number attempted." Since a separate measure is secured for each "fundamental operation" this series of tests may be called diagnostic.

Although this group of tests includes the fundamental operations "represented by currently used textbooks" one may properly raise the question of their relative importance. If we use the results of Douglass's questionnaire, it is obvious that teachers of mathematics do not consider them of equal importance. Some are so lacking in importance according to this criterion that Douglass does not include them even in his list of tests on "additional processes." Furthermore it is well to keep in mind that a battery of sixteen separate tests makes a very elaborate measuring instrument.

Rogers Tests of Mathematical Ability. This group of six tests is designed not to measure achievement but to measure the capacity of pupils to learn algebra and geometry. They differ from general intelligence tests in that their function is more specialized. It is implied that there are certain mental traits which are essential to successful study of elementary algebra and plane geometry. It is these traits which the Rogers Tests of Mathematical Ability are designed to measure. Their function may be described as prognostic. They are intended to be given at the beginning of the school year and the resulting scores used as a basis for advising students in regard to continuing the study of mathematics. If given near the end of the eighth grade the information could be used in advising students entering

high school in regard to undertaking algebra and geometry. The scores are useful also for classifying students in sections to provide for individual differences.

Each of the tests consists of exercises arranged in ascending order of difficulty. They measure power. The first part of *Test 1, Algebraic computation*, consists of exercises like the following:

If $a = 2$, $b = 3$, and $c = 4$, write the values of

$$2a - c, \frac{a + b + c}{c}$$

$3x + 2x + 7x - 4x =$ how many x 's?

If $6x = 30$, what does $x =$?

The second part consists of more difficult exercises including a few simple equations to solve. Students would need to know how to manipulate signs and clear equations of simple fractions.

In *Test 2, Interpolation*, the student is asked to supply the terms in series from which two or more terms have been omitted. In every case the series is formed by the successive additions of a number as in the series, 1, 5, 9, 13, 17, etc. The number added here is 4.

Test 3, Geometry, consists of very simple geometrical theorems. The facts needed in the proof of them are given on the test paper. A preliminary exercise is given so that students who have not studied geometry may understand what they are to do.

Test 4, Superposition, consists of "pairs of symmetrical parallelograms, each with one side on the same straight long black line, and each adjoining a third parallelogram of corresponding design, and similarly with one black edge, but such that it can be superposed upon only one of the adjoining parallelograms. This third parallelogram which has a small circle in one corner is placed in a variety of positions

relative to the pair of parallelograms." The student is asked to indicate which one of the pair of parallelograms the third one will fit if moved only in its plane and to indicate exactly where the circle in the third parallelogram will then lie.

Test 5, Mixed Relations, is similar to an Analogies Tests such as is found in a number of tests of general intelligence. For the last test two of the Trabue completion-Test Language Scales are used. Both Test 5 and Test 6 are verbal and measure certain traits included in general intelligence.

The six tests are published in a single folder and can be administered in about ninety minutes. The scores of the six tests are to be combined in a single score which is used as an index of the student's probable future success in algebra or geometry. In interpreting this index, it must be remembered that such factors as interest, effort, and instruction affect achievement as well as capacity to learn. In addition, these tests as all others are not perfectly reliable. Hence one should not expect complete agreement between the total score yielded by the Rogers Tests of Mathematical Ability and future achievement. The author states that the scores obtained "may be expected to give a correlation of from .60 to .80 with future mathematical achievement." Such coefficients are frequently interpreted as "high," but their departure from 1.00 or perfect correlation is large enough so that one should exercise caution in using the scores yielded by these tests in the educational guidance of students.

Minnick Geometry Tests. The Minnick Geometry Tests were designed to measure separately the following abilities which the author considered important in the demonstration of a theorem in geometry:

1. The ability to draw a figure for the theorem.
2. The ability to state concretely and accurately the hypothesis and conclusion of the theorem

3. The ability to recall additional facts about a figure when one or more facts are given.
4. The ability to select from the available facts those that are necessary for a proof and to arrange them so as to arrive at the desired conclusion.

A separate test was constructed for each of these abilities. The descriptions of the abilities indicate the nature of the exercises. In Test A the student is asked only to draw the figure. In Test B he is given a theorem and a figure to state with reference to the figure "what is given and what is to be proved." The exercises of Test C consist of a figure and a few facts concerning it. The student is asked to "state as many more facts about the figure" as he can. In Test D the student is given a figure and a list of facts about it. He is asked to organize a proof from these facts for a theorem which is stated with reference to the figure. All of the exercises are limited to the first two books of plane geometry. Hence the tests can be given about the middle of the school year.

A student is given two scores: a positive score based upon the number of "necessary statements correctly given" and a negative score which is the "total number of incorrect and unnecessary statements." This dual system of scoring has not been used by other authors of tests. In exercises of the type used in these tests such a negative score is very significant, especially when the tests are considered as instruments for diagnosis.

There are five exercises in Test A, four in Test B and Test C, and three in Test D. Thirty minutes are allowed for each test so no measure of rate is obtained. The exercises are arranged in ascending order of difficulty, but they can be said to form only a very crude scale. Each test contains so few exercises that its reliability is probably very low.

Meeting the teaching situation revealed by algebra tests.

The mental processes of algebra are similar to those of arithmetic in many respects. In both subjects the operations must be performed automatically in order to free the attention for the doing of other things. In arithmetic it was shown that addition involved not a single ability but several abilities. (See page 20.) Each type of example requires a specific ability. A scientific analysis of algebraical abilities has not been made, but it is probable that in algebra each type of exercise requires a specific ability if it is done automatically. The engendering of arithmetical abilities is based upon the laws of habit formation. These laws also apply to the teaching of the operations of algebra. Individual differences complicate the teaching of arithmetic. They have been shown to be equally conspicuous in algebra.

Class instruction possesses the same weaknesses in algebra as it does in arithmetic. The writer has observed the plan of giving drill, described on page 85, in algebra classes as well as in arithmetic. The result was the same in both. Each pupil needs drill upon the types of examples he does not do well. Practice tests for algebra¹ can be devised upon the same principles as those for arithmetic. An interesting series of practice tests in algebra have been devised by Dalman.² The suggestions given on page 83, for adapting the instruction to the needs of the pupils, may be applied to instruction in algebra as well.

Rugg and Clark state that "It has been shown that success in teaching algebra depends primarily on the teacher's knowledge of the typical difficulties which the pupils will meet in learning algebra." In certain reports of the use of

¹ See Rugg, H. O., and Clark, J. R., "Standardized Tests and the Improvement of Teaching in First-Year Algebra"; in *School Review*, vol. 25, pp. 196-213. (March, 1917.)

² Dalman, Murray A., "Hurdles, a Series of Calibrated Objective Tests in First-Year Algebra"; in *Journal of Educational Research*, vol. 1, pp. 47-62. (January, 1920.)

algebra tests the errors made by pupils have been studied. In all cases the number of errors has been large. Table XXV gives the types of errors which were made by two hundred and seventy-five first-year pupils on the Monroe Standard Research Tests in Algebra.¹ The tests were given in March, 1914. Rugg and Clark have used a more elaborate classification of errors, but their results indicate the same condition.

TABLE XXV. CLASSIFICATION OF ERRORS MADE BY TWO HUNDRED AND SEVENTY-FIVE FIRST-YEAR PUPILS ON THE MONROE STANDARD RESEARCH TESTS (ALGEBRA)

TEST	I	II	III	IV	V	VI
Mistake in sign.....	436	265	184	295	387	739
Mistake in the common denominator or in its use.....		167				371
Mistakes in arithmetic.....	143	249	498		596	391
Mistakes in copying.....				382	63	82
One term of binomial not multiplied	29					16
A term neglected.....		103		302	86	26
x omitted.....	117	19		53	80	16
Incomplete as $-x = 5$			38			26

The conditions revealed clearly indicate that the pupils have not been given sufficient satisfactory drill to make automatic the performing of the simpler operations of algebra. Before this can be accomplished, teachers must determine what the important or fundamental operations of algebra are. When this has been accomplished they must further determine what types of exercises occur in each operation. For example, assuming that performing the indicated operation of $a(bx + c)$ is fundamental, it is obvious that these

¹ For a description of these tests see Monroe, Walter S., "A Test of the Attainment of First-Year High-School Students in Algebra"; in *School Review*, vol. 23, pp. 159-71. (March, 1915.)

types of exercises occur, $a(bx + c)$, $-a(bx + c)$, $a(-bx + c)$, $-a(-bx + c)$, $a(bx - c)$, $-a(bx - c)$, $-a(-bx - c)$, $a(-bx - c)$. In the studies referred to it was found that each required its own specific ability. This being the case the teacher must provide satisfactory drill upon each type. A series of scientifically constructed practice exercises furnish a means for doing this.

II. LATIN

Problem of measurement in Latin similar to that in English. In the field of English we have tests of reading, oral and silent, vocabulary, spelling, language, grammar and composition. It would be possible to duplicate these tests in Latin or any other foreign language. The problem of measurement is much the same for one language as for another, but the study of foreign languages in our secondary schools is confined to elementary phases. There is little or no work corresponding to rhetoric and composition in English. Most of the students' time is devoted to "translation," but it involves vocabulary, reading (silent and sometimes oral), and grammar. For this reason most of the tests which have been constructed in Latin and in other foreign languages are designed to measure the ability of students in these three phases of foreign language study.

There is no apparent reason why a silent reading test of the type of the Monroe Standardized Silent Reading Tests, Revised (see page 102) would not be very helpful in teaching Latin or any other foreign language. In fact, a few such tests are beginning to be devised.¹ The extreme emphasis which teachers of foreign languages place upon "translation" rather than "reading" has been criticized. If these critics are right in their contention, silent reading tests in

¹ See the tests by Ullman and Kirby, and Handschin described in the following pages.

foreign languages would serve to direct attention to "reading." No oral reading test has been devised for any foreign language, but a test of the type of the Gray Oral Reading Test (see page 135) would be useful, especially for the modern languages.

The criteria whereby tests in the field of Latin are to be judged are similar to those used in judging the corresponding tests for the English language. The content of the tests should be in agreement with recognized educational objectives. The words for a vocabulary test should be representative of the vocabulary which students meet in their study of the language. The sentences of a translation or reading test should be representative of the texts which they study. The tests should be highly objective with respect to the scoring of the test papers. Reliability is also an important criterion.

It should be recognized that Latin and the other foreign languages taught in the secondary schools afford greater opportunities for the use of the results of measurement in improving instruction than is found in any other subject. As we pointed out on page 300 there is little opportunity for diagnosis in subjects where the student is engaged in studying one topic after another. In a foreign language a student is expected to continue enlarging his vocabulary, his knowledge of grammar, and his ability to read and translate. Thus there is abundant opportunity to make and use diagnostic measurements.

Godsey Diagnostic Latin Composition Test. This test is designed to measure ability of pupils to translate simple English sentences into Latin and to indicate the grammatical rule which determines the form of one word in the Latin sentence. The exercise consists of an English sentence and its translation into Latin. In this translation one word is given in four forms. The student is directed to

draw a circle around the correct form. Two sample exercises are given below.

- a. The king's sons are leaders of the enemy.

Regis (filius, filiorum, filios, filii) duces hostium sunt.....
4..7..13..14

- b. They were hurling weapons at the wall.

Tela in murum (coniciebantur, coniciebant, coniciunt, conicerent).....1..8..13..14

The numbers at the end of the Latin sentence refer to grammatical rules which are printed at the bottom of the page of the test folder. The pupil is directed to "draw a circle around the number of the rule which applies to the correct form." The test consists of thirty-three exercises. Students are allowed thirty minutes for doing them.

Henmon Latin Tests. Henmon has prepared two series of Latin tests, one for measuring the student's acquaintance with Latin vocabulary and the other for measuring his ability to translate Latin sentences into English. In order to determine the words to be included in his vocabulary test Henmon tabulated all of the Latin words occurring in thirteen widely used first-year texts. Three hundred and nineteen words were found common to all of these books. This list was compared with Lodge's list of words used in Cæsar, Cicero, and Virgil. Those words in the original list were excluded which were not used by all three of these authors. The final list consisted of 239 Latin words which were common to thirteen widely used first-year texts and also to Cæsar, Cicero, and Virgil. This list Henmon calls a "standard vocabulary for high-school Latin."

Henmon submitted these words to high-school students with a request that they write what they considered the English equivalent. From the data collected in this way the per cent of correct responses was calculated for each word. Four vocabulary tests of fifty words each were con-

structed so that they are essentially duplicate forms. In each test the words are arranged in ascending order of difficulty. For measuring ability of students to translate Latin into English, Henmon composed Latin sentences in which only words occurring in his fundamental vocabulary list appeared. The difficulty of these sentences has been determined and they are arranged in ascending order of difficulty. The student is asked to write his translation in English.

Holtz-Godsey Latin Teaching Tests. The five tests of this series relate to Latin vocabulary and are intended to measure certain of the indirect outcomes generally supposed to result from the teaching of Latin. Test A and Test B are designed to measure the pupil's acquaintance with English derivatives of Latin words. In Test A the student is given a list of Latin words and asked to write as many English derivatives as he can for each. In Test B he is given a list of English words which are derived from the Latin and a list of the Latin words from which the first list is derived. The English words are numbered and the student is directed to "number the Latin source word with the number of its English derivatives."

Test C consists of a list of English words which are derived from the Latin. The student is asked to define these words and to use them in sentences. Test D also consists of a list of English words derived from Latin. Some of these words are incorrectly spelled and the student is asked to write the correct spelling for these words. He is also asked to write the plurals of certain words. These words are taken from the Latin and their plural in most cases is the Latin plural. Test E consists of Latin words arranged in the form of a "same-opposite" test.

As the title of these tests indicates, their purpose is primarily instructional. They should be helpful to teachers

who are attempting to engender the indirect outcomes which they are designed to measure.

Pressey Test in Latin Syntax (nouns, pronouns, and adjectives). The exercises of this test consist of an English sentence and four Latin sentences. One of the Latin sentences is a correct translation of the English sentence. The student is directed to draw a line under the one Latin sentence which is a correct translation of the English sentence. Two sample exercises are given below.

1. He killed the soldier with a sword.

Militem cum gladiō interfēcit. Militem gladiō interfēcit.
Militem ā gladiō interfēcit. Militem gladium interfēcit.

2. Cæsar drew up a triple line of battle.

Cæsar aciem triples instrūxit. Cæsar aciem triplicam instrūxit.
Cæsar aciem triplicem instrūxit. Cæsar aciem triplici instrūxit.

The test consists of thirty-three exercises of this kind arranged in ascending order of difficulty. The student is allowed twenty minutes for the test.

Tyler-Pressey Test in Latin Verb Forms. This test is designed to measure the student's understanding of Latin verb forms. It consists of exercises in which a Latin verb form is given together with four English translations. The student is directed to underline the translation which he considers correct. There are thirty-two exercises and the time allowance is fifteen minutes. A sample exercise is given below.

1. Movemus

We shall move . . . Let us move . . . We are moved . . . We are moving . . .

Starch-Waters Latin Tests. The Starch-Waters Latin vocabulary test consists of one hundred words which were selected at random from Lodge's Vocabulary of High School

Latin. This vocabulary list contains the two thousand words occurring in Cæsar, Cicero, and Virgil. In the test the words are arranged in alphabetical order and the student is asked to write the English meaning of the Latin words.

The translation test consists of four parts, one for each of the four years of Latin taught in the high school. In the case of each test the sentences were selected so as to form a random sample of the translation required of students. The test for the first year consists of twenty sentences; those for the second and third years of seven each; and the test for the fourth year of ten sentences. The student is directed to write the translation just below each line of Latin. "If you know the meaning of a word but do not know its use, write the English word in parentheses, in the place where the English translation of that word should be written. If you do not know the meaning of a word but know its construction write the construction in parentheses, e.g. (gen. of possession) in the place where the English translation of that word should be placed."

From our experience in scoring silent reading tests in which the pupil is asked to reproduce from memory the material read, and also the answers to ordinary questions one would expect to find the scoring of this translation test highly subjective. Even the vocabulary test is probably sufficiently subjective to affect seriously its reliability. In the absence of any determination of reliability one would expect both of these tests, particularly the one on translation, to rank very low in this characteristic.

Ullman-Kirby Latin Comprehension Test. This is a Latin silent reading test very similar to the Thorndike-McCall Reading Scale. (See page 120.) The student is given a paragraph of Latin to read and then he is asked certain questions concerning it. These questions are asked in English and the student is expected to answer them in

English. There are ten exercises arranged in ascending order of difficulty. Thirty minutes are allowed for the test. Exercise 2 is given below.

Vir et filius cum servō ad urbem Rōmam iter faciēbant. Primā lūce profecti erant. Ad flūmen altum vēnērunt quod primō trānsire nōn poterant. Sed nāvis parva ā servō vīsa est in quā duōbus sōlis locus erat. Itaque nāvī servus cum puerō transiit; tum patrem pueri trāsportāvit. Eādem nocte ante primam lūcem omnēs ad urbem pervēnērunt.

5. About how long did it take the travelers to reach Rome?

.....

6. Who found the boat?.....

7. Who were in the boat on the last trip across the river?

.....

From our experience with silent reading tests of this type one would expect that the scoring of the test would not be as objective as one in which the student was directed to underline the answers which he considers correct. However, this test is probably as highly objective as the Thorndike-McCall Reading Scale and other tests of this type.

III. MODERN LANGUAGES

Problem of measurement. The problem of measurement in modern languages is very similar to that in Latin. The most significant difference is the greater need for a test of ability in oral reading. For this purpose a test similar to the Gray Oral Reading Test would prove very helpful. However, no such test has been constructed. The number of tests for any of the modern languages is considerably smaller than the number for Latin. The apparent reason for this is the greater interest of teachers of Latin in certain phases of the teaching of their subject and the common content of the courses in Latin.

Handschin Silent Reading Tests in Spanish. Test A con-

sists of fourteen exercises printed in Spanish. They are somewhat similar to the exercises of the Kansas Silent Reading Test.¹ The pupil is required to write his answers in Spanish. Some of these are numbers, but they must be expressed in words. In general, one and only one answer is correct which makes the test highly objective in the scoring. Five minutes are allowed for the test.

Test B consists of a paragraph in Spanish which the pupils are directed to read "as rapidly as possible but be sure to get the meaning." The pupils are also told that they will be asked to answer questions, and they are given an opportunity to see the questions before the paragraph is read, but they are not permitted to consult the questions during the time of reading or the paragraph when they are writing the answers to the questions. One minute is given for reading a paragraph of 155 words. Five minutes are given for answering ten questions. The pupils are directed to draw a circle around the last word read when time is called. This provides for a rate score. A comprehension score is secured from the number of questions which are answered correctly.

Handschin Silent Reading Tests for French. These are very similar to the two corresponding tests for Spanish. In Test B the paragraph to be read is a little longer, but this is the principal difference. Both of the tests are designed to be used in either the first or second year of the study of this language.

Handschin Comprehension and Grammar Test A — French. This test is designed for students in their first year of the study of French. It contains six easy French sentences which are to be studied for five minutes for the purpose of reproducing them correctly in French. The sentences have certain words omitted. The student is expected

¹ See Monroe Standardized Silent Reading Tests, page 99.

to supply the missing words in French. After the sentences have been reproduced then the student is directed to re-write them "in the third person, plural, past indefinite tense," changing all words to the proper form. The student is given a score for both versions of the sentences.

Henmon French Tests. These tests are very similar to the ones devised by Henmon to measure ability in Latin. Each test consists of a vocabulary list of sixty words and twelve sentences to be translated. In both cases the exercises are arranged in ascending order of difficulty. In addition to measuring the scope and accuracy of vocabulary and ability to understand connected discourse in French the tests are intended to measure knowledge of grammar. The words appearing in these tests are among those which are found to be common to "twelve recent and widely used first-year texts." The words found in the sentences also occur in this list of common words.

There are four tests, which may be used as duplicate forms. The tests are intended to measure a pupil's power rather than rate of work. The sum of the difficulty values of the words whose meaning is given correctly constitutes the pupil's score on the vocabulary test. The pupil's score on the translation of the sentences is the sum of the difficulty values of the sentences translated correctly. Eight minutes are allowed for the vocabulary test and twelve for the sentences.

Wilkins Prognosis Test in Modern Languages. The Wilkins Prognosis Test in Modern Languages is designed to determine the probable ability of students to learn a foreign language. It is expected that this information will be used for eliminating from modern language courses those pupils who will probably be unsuccessful and for classifying the students who can learn the language on the basis of their apparent abilities. The test may be given on the first day

of the term and should in every case be administered before the students "have been taught anything at all of the foreign language." Provision is made for an "elimination test" in the particular foreign language studied (French or Spanish) at the end of four weeks of study of the language. The prognostic test consists of six parts.

I. Visual-motor (seeing and writing). A flash card on which appears a short sentence in either French or Spanish is exposed for five seconds. The pupils are directed to observe each card and then in ten seconds write the sentence which they saw.

II. Aural-motor (seeing and hearing). In this test sentences (Spanish or French) are pronounced to the class. After a sentence is pronounced three times the pupils are directed to write it.

III. Memory. In this test the pupils are given a list of ten words taken from a foreign language and the English equivalent of each. They study this list for ten minutes and then attempt to write from memory the English equivalent for each foreign word.

IV. Grammar concepts. The students are given English sentences and asked to make certain changes in the word forms. For example, sentences in the present tense or in the past tense are to be changed to future time. Sentences which have a singular subject are to be changed so as to have a plural subject.

The remaining two tests of the series are to be given to pupils individually. *Test V — Visual-oral* (seeing and speaking) is in English. It is intended to measure a pupil's ability to read a sentence silently and then repeat it orally from memory. *Test VI — Aural-oral* (hearing and speaking) consists of ten sentences in the foreign language (French or Spanish) which are read to the pupils one at a time. Each sentence is read three times and after the third reading the pupil is asked to repeat the sentence orally.

The pupil's composite score on this group of six tests is taken as an index of his ability to learn a modern language.

IV. SCIENCE

Problem of measurement. The problem of measurement in any of the sciences taught in the high school is made difficult by reason of the fact that there has been no authoritative determination of minimum essentials of each subject. This is a prerequisite step for the construction of an achievement test. An additional difficulty is created by reason of the fact that the subjects are divided into rather distinct topics. This is especially true in the case of physics. A single general test to be given at the end of the year would be unsatisfactory. Unless there has been an intensive review students would be at a disadvantage in taking a test on topics studied early in the year.

Prognostic tests such as Van Wageningen's Reading Scale for General Science should be very useful. Such a test could be used to select those students who are unlikely to succeed in the study of a science. Information tests designed to measure acquaintance with technical terms should also prove useful.

Downing Range-of-Information Test in Science. This test consists of a list of fifty technical terms from the field of science arranged in alphabetical order. The student is asked to put an "E" beside the words which he can explain or define, an "F" beside those he has heard of or read about but for which he does not have a clear meaning, and an "N" beside those which are new. He is then asked to explain or define the first five which are marked "E." The words are about equally divided between the following three fields: (1) biology, (2) physiology and geography, (3) physics and chemistry. The type of response required makes the test highly subjective because all students will not interpret the

directions in the same way. A better test could be made by requiring a response similar to that called for by the Thorndike Test of Word Knowledge. (See page 132.) A longer list would make a more reliable test.

Starch Physics Test. Starch has devised a series of tests in physics, covering mechanics, heat, sound, light and magnetism and electricity. The tests consist of sentences, from which words have been omitted. The sentences and the words to be omitted have been chosen so that a pupil cannot supply the correct words unless he knows certain physical facts or principles. The following exercises will illustrate the nature of the tests:

17. The periods of pendulums of equal lengths swinging through short arcs are independent of.....
and also independent of.....
30. The point.....degrees below.....
degrees centigrade is called.....
39. The frequency of vibration of a string varies inversely as
53. The critical angle is that angle of incidence which will produce
55. Electromotive force is the difference in.....
between.....

The tests consist of seventy-five mutilated sentences of this type. The facts, principles, and laws upon which these sentences are based were determined by examining five widely used textbooks. The one hundred and two facts, principles, or laws which were treated by all five of the textbooks are the ones which the pupils must know to do the exercises correctly. Starch considered these facts and principles probably to be the ones fundamental to elementary physics. If the tests do nothing more than define the fundamental facts and principles they will fulfill an important

function. However, they also give the teacher a means for comparing his class with other classes.

Iowa Physics Test. This series of tests was devised by H. L. Camp and includes a test for each of the following divisions of physics: (1) mechanics, (2) heat, (3) electricity and magnetism. Some of the exercises are mathematical problems to be solved, others are questions calling for certain specific facts. The exercises have been arranged in ascending order of difficulty. The pupil's score is the sum of difficulty values of those exercises done correctly.

Van Wagenen Reading Scale for General Science. This test is designed to measure a pupil's ability to read paragraphs which are typical of science texts, particularly in the field of general science. The exercises consist of a paragraph to be read and a series of statements. The pupil is asked to check each statement which "contains an idea that is in the paragraph or can be derived from it." The exercises are arranged in ascending order of difficulty and the pupil's score is the degree of difficulty of the most difficult exercise which he does with a specified degree of success. The procedure to be followed in computing the score is difficult to understand and will be confusing to many teachers although tables are provided to facilitate the work. This test may be described as prognostic. It may be given at the beginning of a course and should indicate those students who are not qualified to undertake the study of science.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. How do you account for the fact that less progress has been made in devising tests for high-school subjects than for elementary-school subjects?
2. Which of the algebra tests described in this chapter do you think would be most helpful to a teacher? Why?
3. Compare the Latin tests which are described. Which one do you think would be most helpful to a teacher? Why?

4. Is a test like the physics tests superior to an ordinary examination? Justify your answer.
5. The Starch Physics Test covers a very wide range of topics. Is this necessary for a test in such subjects as geometry, physics, or history?
6. If you are teaching algebra, study the errors which your pupils make. Suggest a plan for reducing the number of errors.
7. Suggest several plans for increasing the foreign language vocabulary of pupils who have been shown to be below standard. How could you determine which of the plans is the best?
8. What should be the attitude of high-school teachers toward standardized achievement tests? Why?
9. Are there any satisfactory standardized tests in the field of science? Is it likely we will ever have any comparable to the tests we now have in silent reading? Why?

SELECTED BIBLIOGRAPHY

1. MATHEMATICS

- Cawl, Franklin R. "Practical Uses of an Algebra Standard Scale"; in *School and Society*, vol. 10, pp. 88-90. (July 19, 1919.)
- Courtis, S. A. "The Measurement of High-School Mathematics"; in *School Science and Mathematics*, vol. 18, pp. 507-26. (June, 1918.)
- Dalman, Murray A. "Hurdles, a Series of Calibrated Objective Tests in First-Year Algebra"; in *Journal of Educational Research*, vol. 1, pp. 47-62. (January, 1920.)
- Douglass, Harl Roy. *The Derivation and Standardization of a Series of Diagnostic Tests for the Fundamentals of First-Year Algebra*. University of Oregon Publication, vol. 1, no. 8. (Eugene: University of Oregon, 1921. 48 pp.)
- Douglass, Harl Roy. "A Series of Standardized Diagnostic Tests in the Fundamentals of Elementary Algebra"; in *Journal of Educational Research*, vol. 4, pp. 396-403. (December, 1921.)
- Harris, Eleanora, and Breed, Frederick A. "Comparative Validity of the Hotz Scales and the Rugg-Clark Tests in Algebra"; in *Journal of Educational Research*, vol. 6, pp. 393-411. (December, 1922.)
- Hobbs, James B. "Results from Giving the Hotz First-Year Algebra Scale Test to a Six-Eight Months Group"; in *School and Society*, vol. 12, pp. 353-54. (October 16, 1920.)
- Hotz, G. H. *First-Year Algebra Scales*. Teachers College Contributions to Education no. 90. (New York: Teachers College, Columbia University, 1918.)
- Irwin, H. N. "Preliminary Attempt to Devise a Test of the Ability of High-School Pupils in the Mental Manipulations of Space Relations"; in *School Review*, vol. 26, pp. 600-06, 654-70, 759-72. (October, December, 1918.)

- Mensenkamp, L. E. "Tests of Mathematical Ability and their Prognostic Values; a Discussion of the Rogers Tests"; in *School Science and Mathematics*, vol. 21, pp. 150-62. (February, 1921.)
- Miller, G. A. "Analysis of Mathematical Abilities"; in *School and Society*, vol. 7, pp. 683-84. (June 8, 1918.)
- Minnick, J. H. "Certain Abilities Fundamental to the Study of Geometry"; in *Journal of Educational Psychology*, vol. 9, pp. 83-90. (February, 1918.)
- Minnick, J. H. "A Scale for Measuring Pupils' Ability to Demonstrate Geometrical Theorems"; in *School Review*, vol. 27, pp. 101-09. (February, 1919.)
- Minnick, J. H. "The Scoring of Geometry Test W"; in *Educational Administration and Supervision*, vol. 6, pp. 509-11. (December, 1920.)
- Monroe, Walter S. "A Test of the Attainment of First-Year High-School Students in Algebra"; in *School Review*, vol. 23, pp. 159-71. (March, 1915.)
- Monroe, Walter S. "Some Correlations between Otis Scale and Rogers Mathematical Tests"; in *Journal of Educational Research*, vol. 2, pp. 774-76. (November, 1920.)
- Monroe, Walter S. "Measurement of Certain Algebraic Abilities"; in *School and Society*, vol. 1, pp. 393-95. (March 13, 1915.)
- Rogers, Agnes L. *Experimental Tests of Mathematical Ability and their Prognostic Value*. Teachers College Contributions to Education no. 89. (New York: Teachers College, Columbia University, 1918.)
- Rugg, H. O. "The Experimental Determination of Standards in the First-Year Algebra"; in *School Review*, vol. 24, pp. 37-66. (January, 1916.)
- Rugg, H. O., and Clark, J. R. "The Improvement in Ability in the Use of the Formal Operations of Algebra by Means of Formal Practice Exercises"; in *School Review*, vol. 25, pp. 546-54. (October, 1917.)
- Rugg, H. O., and Clark, J. R. "Standardized Tests and the Improvement of Teaching in First-Year Algebra"; in *School Review*, vol. 25, pp. 113-32, 196-213, 346-49. (February, March, May, 1917.)
- Stockard, L. V., and Bell, J. C. "A Preliminary Study of the Measurement of Abilities in Geometry"; in *Journal of Educational Psychology*, vol. 7, pp. 567-80. (December, 1916.)
- Thorndike, Edward L. "Instruments for Measuring the Disciplinary Values of Studies"; in *Journal of Educational Research*, vol. 5, pp. 269-79. (April, 1922.)
- Williams, Lewis W. "Illinois Standardized Algebra Test"; in *Journal of Educational Research*, vol. 3, pp. 75-76. (January, 1921.)

2. LATIN

- Brown, H. A. *Latin in Secondary Schools*. (Oshkosh, Wisconsin: State Normal School, 1919. 170 pp.)
- Carr, W. L., and Gray, Mason D. *Preliminary Report Submitted by the In-*

investigators to the National Committee on the Teaching of Latin and Greek in the Secondary Schools of the United States. (1921. 43 pp.)

Handschin, C. H. "A Test for Discovering Types of Learners in Language Study"; in *Modern Language Journal*, vol. 3, pp. 1-4. (October, 1918.)

Hanus, Paul R. "Measuring Progress in Learning Latin"; in *School Review*, vol. 24, pp. 342-45. (May, 1916.)

Henmon, V. A. C. "The Measurement of Ability in Latin, Part I, Vocabulary; Part II, Sentence Tests"; in *Journal of Educational Psychology*, vol. 8, pp. 515-38, 589-99; vol. 11; pp. 131-36. (November, December, 1917, March, 1920.)

Hobson, E. G. "Observations on Two Latin Vocabulary Tests"; in *School Review*, vol. 28, pp. 509-17. (September, 1920.)

Lohr, Lawrence L. "A Latin Form Test"; in *The High School Journal*, vol. 5, pp. 217-23. (December, 1922.)

Starch, Daniel. "A Test in Latin"; in *Journal of Educational Psychology*, vol. 10, pp. 489-500. (December, 1919.)

3. MODERN LANGUAGE

Briggs, Thomas H. "Prognosis Tests of Ability to Learn Foreign Languages"; in *Journal of Educational Research*, vol. 6, pp. 386-92. (December, 1922.)

Davidson, Edna H. "A Report on the Use of the Handschin Modern Language Tests in the Jamaica High School"; in *Bulletin of High Points*, vol. 2, pp. 6-7. (November, 1920.)

Handschin, Charles H. "A Test for Discovering Types of Learners in Language Study"; in *Modern Language Journal*, vol. 3, pp. 1-4. (October, 1918.)

Mahler, Miriam. *A Report on the Wilkins Prognosis Test in French Given in Public School 91, Manhattan.* Bulletin of High Points in the Work of the High Schools of New York City, vol. 3, no. 5, p. 14. (May, 1921.)

Wilkins, L. A. *Testing for Ability to Learn a Foreign Language.* Bulletin of High Points in the Work of the High Schools of New York City, vol. 1, no. 2, p. 5 (February, 1919); vol. 1, no. 8, p. 26. (October, 1919.)

4. SCIENCE

Bell, J. Carleton. "A Test in First-Year Chemistry"; in *Journal of Educational Psychology*, vol. 9, pp. 199-209. (April, 1918.)

Bell, J. Carleton. "Study of the Attainments of High-School Pupils in First-Year Chemistry"; in *School Science and Mathematics*, vol. 18, pp. 425-32. (May, 1918.)

Briggs, Thomas H. "Results of the Bell Chemistry Test"; in *Journal of Educational Psychology*, vol. 11, pp. 224-28. (April, 1920.)

Camp, Harold L. "Scales for Measuring Results of Physics Teaching"; in *Journal of Educational Research*, vol. 5, pp. 400-05. (May, 1922.)

- Chapman, J. Crosby. "The Measurement of Physics Information"; in *School Review*, vol. 27, p. 748. (December, 1919.)
- Downing, E. R. "Additional Science Tests in the Grades"; in *Nature Study Review*, vol. 16, pp. 62-65. (February, 1920.)
- Downing, E. R. "A Range of Information Tests in Science"; in *School Science and Mathematics*, vol. 19, pp. 228-33. (March, 1919.)
- Gerry, Henry L. "Further Data on the Bell Chemistry Test"; in *Journal of Educational Psychology*, vol. 11, pp. 398-401. (October, 1920.)
- Grier, N. M. "Range of Information Test in Biology"; in *Journal of Educational Psychology*, vol. 9, pp. 210-16, 388-93. (April, September, 1918.)
- Grier, N. M. "The Range of Information in Biology; III, Botany"; in *Journal of Educational Psychology*, vol. 10, pp. 509-16. (December, 1919.)
- Hayes, Seth. "Coöperative Chemistry Tests"; in *Journal of Educational Research*, vol. 4, pp. 109-20. (September, 1921.)
- Jones, F. T. "Practice Exercises in Physics"; in *School Review*, vol. 26, pp. 341-48. (May, 1918.)
- Maxwell, P. A. "Standardization of 'First-Year Science Tests'"; in *General Science Quarterly*, vol. 5, pp. 226-31. (May, 1921.)
- Powers, S. R. "A Comparison of the Achievement of High-School and University Students in Certain Tasks in Chemistry"; in *Journal of Educational Research*, vol. 6, pp. 332-43. (November, 1922.)
- Randall, D. P., Chapman, J. C., and Sutton, C. E. "The Place of the Numerical Problem in High-School Physics"; in *School Review*, vol. 26; pp. 39-43. (January, 1918.)
- Ruch, G. M. "A Range of Information Test in General Science"; in *General Science Quarterly*, vol. 4, pp. 257-62. (November, 1919.)
- Ruch, G. M. "Range of Information Test in General Science: Preliminary Data on Standards"; in *General Science Quarterly*, vol. 5, pp. 15-19. (November, 1920.)
- Starch, Daniel. "The Measurement of Efficiency in Spelling, and the Overlapping of Grades in Combined Measurements of Reading, Writing, and Spelling"; in *Journal of Educational Psychology*, vol. 6, pp. 167-86. (March, 1915.)
- Van Wagenen, M. J. "Van Wagenen Reading Scales in History, General Science, and English Literature"; in *Journal of Educational Research*, vol. 3, pp. 314-16. (April, 1921.)
- Webb, Hanor A. "A Preliminary Test in Chemistry"; in *Journal of Educational Psychology*, vol. 10, pp. 36-43. (January, 1919.)

5. OTHER SUBJECTS

- Blackstone, E. G. "Measurement of Progress in Typewriting"; in *Detroit Journal of Education*, vol. 1, pp. 35-41. (May, 1921.)
- Blackstone, E. G. "Tentative Standards in Typewriting"; in *Detroit Journal of Education*, vol. 1, p. 53. (December, 1920.)

- Briggs, Thomas H. "A Dictionary Test"; in *Teachers College Record*, vol. 24, pp. 355-65. (September, 1923.)
- Brown, Clara M. "Investigations Concerning the Murdoch Sewing Scale"; in *Teachers College Record*, vol. 23, pp. 459-70. (November, 1922.)
- Chassell, Clara F., Upton, Siegfried Maia, and Chassell, Laura M. "Short Scales for Measuring Habits of Good Citizenship"; in *Teachers College Record*, vol. 23, pp. 52-79. (January, 1922.)
- Cody, Sherwin. *Commercial Tests and How to Use Them*. (Yonkers, New York: World Book Company, 1919.)
- Gaw, Esther Allen. "Some Individual Difficulties in the Study of Music"; in *Journal of Educational Research*, vol. 5, pp. 381-88. (May, 1922.)
- Hoke, Elmer Rhodes. *The Measurement of Achievement in Shorthand*. The Johns Hopkins University Studies in Education no. 6. (Baltimore: The Johns Hopkins Press, 1922. 118 pp.)
- King, Florence B. *A Measuring Scale in Foods*. Indiana University Extension Division Bulletin, vol. 7, no. 10, pp. 144-46. (Bloomington: University of Indiana, 1922.)
- Murdoch, Katharine. "A New Analytic Sewing Scale"; in *Teachers College Record*, vol. 23, pp. 453-58. (November, 1922.)
- Murdoch, Katharine. *The Measurement of Certain Elements of Hand Sewing*. Teachers College Contributions to Education no. 103. (New York: Teachers College, Columbia University, 1919. 120 pp.)
- Rugg, H. O. "A Scale for Measuring Free-Hand Lettering"; in *Journal of Educational Psychology*, vol. 6, pp. 25-42. (January, 1915.)
- Seashore, Carl E. *A Survey of Musical Talent in the Public Schools*. University of Iowa Studies in Child Welfare, vol. 1, no. 2. (Iowa City: University of Iowa, 1920. 36 pp.)
- Seashore, Carl E. *The Psychology of Musical Talent*. (Boston: Silver, Burdette and Company, 1919. 288 pp.)
- Seashore, Carl E. "The Measurement of Pitch Discrimination: a Preliminary Report"; in *Journal of Educational Psychology*, vol. 2, p. 278. (1911.)
- Seashore, Carl E. "Measurement of Musical Talent"; in *Musical Quarterly*, vol. 1, pp. 129-48. (January, 1915.)
- Seashore, Carl E. *Manual of Instructions and Interpretations for Measures of Musical Talent*. (New York: Columbia Graphophone Company, Educational Department, 1919. 16 pp.)
- Trilling, Mabel Barbara, and Hess, Adah. "Informal Tests in Teaching Textiles and Clothing"; in *Journal of Home Economics*, vol. 13, pp. 483-89. (October, 1921.)
- Trilling, Mabel Barbara, Miller, Ethelwyn, and others. Chapter VII — "Measuring the Results of Teaching in Textiles, Dress Design, Sewing, and House Planning"; chapter VIII — "Scales for Measuring Skill in Machine Sewing." *Supplementary Educational Monographs*, vol. 2, no. 6, pp. 75-114. (Chicago: University of Chicago, 1920.)

- Tuttle, W. W. "The Determination of Ability for Learning Typewriting"; in *Journal of Educational Psychology*, vol. 14, pp. 177-81. (March, 1923.)
- "A Measuring Scale for Gregg Shorthand Penmanship"; in *Journal of Educational Research*, vol. 8, pp. 79-80. (June, 1923.)
- "Tests in Home Economics"; in *Proceedings of the High School Conference of November 23, 24, and 25, 1922*, pp. 249-54. (Urbana: University of Illinois, 1923.)

CHAPTER IX

INTELLIGENCE TESTS

I. THE PROBLEM OF THE MEASUREMENT OF GENERAL INTELLIGENCE

Measurement of general mental ability not new. The attempt to measure the general mental ability of an individual is neither new nor revolutionary. Men have always passed judgment on their fellows in such terms as "Jones is more capable than Brown," or, "John is only half as bright as Roy." Both feeble-mindedness and genius were recognized types of mental endowment long before the mental tests were invented. The modern testing movement is best understood when seen as an attempt to refine and make precise estimates which have been necessary as long as mankind has tried to live together in social groups.

Definition of intelligence. The word "intelligence" has been a handicap to the measurement movement. Like many other terms with long-established usage which psychology has adopted, it has several well-recognized meanings. Because it has been associated with genius, feeble-mindedness, and insanity, it carries the taboo of personal and intimate things. Because it has been the chief characteristic which distinguished man from the other animals, and because it has frequently stood as an attribute of Deity, it has a sacred and spiritual significance which leads some to resent its measurement and causes others to regard it as the sum total of man's mind and soul. In order to think clearly of the results of intelligence tests, we must relieve the concept of some of its handicaps.

An eclectic view of intelligence. Educational practice

demands a prediction of the pupil's probable school success. The word "intelligence" is generally accepted as a name for such a prediction. Colvin¹ has said, "Therefore, to identify general intelligence with native learning ability is, both theoretically and practically, justifiable." Later Monroe² states, "For practical school purposes general intelligence may be thought of as the measurement of the pupil's general capacity to do the work of the school." These definitions point out the purpose of intelligence measurement as it is treated in this chapter. They do not dispose of certain issues which are frequently raised.

One of the chief of these issues is the native or innate character of intelligence. Is intelligence independent of learning? For ordinary school practice this issue may be ignored. The typical teacher faces children who have had much opportunity to learn. The abilities which they bring with them to the classroom are doubtless resultants of both environmental and innate forces. To assume that the general ability of a child cannot be modified would tend to discourage all attempt to educate. On the other hand, to assume that a feeble-minded child can be prepared for high school or college leads to waste effort and to punishment of the child. A safe assumption seems to be that children's general intelligence is natively determined within such limits that they can be safely grouped on this basis for instructional purposes. With such grouping it is certain that an opportunity should exist for children who have special abilities. Whether general intelligence is ultimately determined to be due wholly to innate or partly to innate factors, it is certain that our meas-

¹ Colvin, S. S., "Principles Underlying the Construction and Use of Intelligence Tests"; in *Twenty-First Yearbook of the National Society for the Study of Education*, Part I, p. 17. (Bloomington, Illinois: Public School Publishing Company, 1922.)

² Monroe, Walter S., *Introduction to the Theory of Educational Measurements*, p. 39. (Boston: Houghton Mifflin Company, 1923.)

urements are still so unreliable that we must regard the estimated intelligence of a child with a suspicion that it may be higher. This attitude leads to the same procedure as that which we would follow if we believed that general intelligence may be improved.

The mention of general and special abilities in the same paragraph suggests the issue raised by those who hold that intelligence is a general ability, or the "general common factor" of the English psychologists. The opposing view is that it is a composite of many distinctive abilities. Whatever it may be, we can say that now it is being measured by a variety of tests which appear to involve different abilities, but which when used to secure a composite score give a good prediction of a child's school success.

Both of these issues are involved in the consideration of the nature of intelligence. Since these questions, as well as some others, are unsettled, some insist that we cannot measure that which we cannot describe. This is a question for the metaphysician. While waiting for the final word we have plenty of precedent for proceeding with measurement. Although we are ignorant of the true nature of intelligence, we are also ignorant of the true nature of light, heat, time, and electricity, but in some way we measure each of these things.

General plan of measurement. General intelligence or capacity to learn is measured indirectly by measuring what a child has learned. In order that this indirect measurement may have maximum validity, the opportunities for acquiring the achievements measured should be as nearly the same as possible for all children tested. Several of the general intelligence tests described in this chapter measure certain achievements in language, arithmetic, and general information. Other tests measure abilities to do things which are not taught in school. Examples of tests of this sort are

"directions tests," and "substitution tests." In such cases the assumption is made that, since no child has been definitely taught to do the things which the test calls for, all will be tested on the same basis.

Measures of achievement secured by means of standardized tests in language, reading, or arithmetic have been used as measures of general intelligence. The Trabue Language Completion Test is now used much more frequently for measuring general intelligence than for measuring achievement in the field of language. Vocabulary tests have also been used as intelligence tests.

Certain of the individual intelligence tests yield scores in terms of mental age, but most of the tests described in the following pages yield point scores. These may be translated into mental ages. Tables or specific directions for doing this may be found in the manuals of directions which accompany the tests. A pupil's mental age describes his level of intelligence. The quotient obtained by dividing the pupil's mental age by his chronological age is called his intelligence quotient (I.Q.). This measure is an index of the pupil's brightness, and furnishes a basis for predicting the level of intelligence which he will probably reach at some future date.

II. DESCRIPTIONS OF REPRESENTATIVE INTELLIGENCE TESTS

The large number of available tests for measuring intelligence makes it impossible to describe in a single chapter more than a few of the most widely used tests. These, however, may be considered to be representative of the larger group of available tests. For descriptions of tests not given here the reader is referred to current bibliographies.¹ For conven-

¹ Whipple, G. M., "An Annotated List of Group Intelligence Tests"; in *Twenty-First Yearbook of the National Society for the Study of Education*,

ience the tests have been grouped for purposes of description under the following heads:

1. Tests of the Binet-Simon type.
2. Individual performance tests.
3. Group tests for literate school children.
4. Non-verbal intelligence tests.
5. Group tests for college students.

The last three of these groups overlap in the sense that the tests described in one group may occasionally be used for measuring the intelligence of persons indicated in another group. However, this classification will assist the reader in making comparisons between tests designed for similar purposes.

1. Tests of the Binet-Simon type

Tests of this type may be said to be revisions¹ of the original Binet-Simon tests. Binet prepared the first edition Part I, pp. 93-113. (Bloomington, Illinois: Public School Publishing Company, 1922.)

Boardman, Helen, *Psychological Tests — A Bibliography*. Bureau of Educational Experiments, Bulletin 6. (New York: Bureau of Educational Experiments, 1918. 111 pp.)

¹ Revisions of the Binet Scale consist in (1) relocating the tests on the scale, (2) eliminating certain tests, and (3) adding certain tests. For example, Terman in his *The Measurement of Intelligence*, page 61, compares the Stanford Revision with the Binet Scale. "On the whole it differs somewhat more from the Binet 1911 scale than from that of 1908. Thus, of the 49 tests below the 'adult' group in the 1911 scale, 2 are eliminated and 29 are relocated. Of these, 25 are moved downward and 4 upward. The shifts are as follows:

Down 1 year, 18
Down 2 years, 4
Down 3 years, 2
Down 6 years, 1
Up 1 year, 3
Up 2 years, 1

"Of the adult group in Binet's 1911 series 1 is eliminated, 2 are moved up to 'superior adult,' and 1 is moved up to 14. Accordingly, of Binet's entire

in 1905. The 1908 and 1911 editions were prepared in collaboration with Dr. Simon. Revisions appeared rapidly in many lands. Decroly and Degand published a revision in Belgium in 1910, Goddard in America and Johnston in England in 1911, and Bobertag in Germany in 1913.

The Stanford Revision. The most popular form of this scale in this country is the Stanford Revision, prepared by Terman in 1916.¹ The general procedure in giving this test is representative of that of the other revisions mentioned. It is made up of a set of six regular exercises, with one or more alternates for each year from three to ten inclusive. For year twelve there are eight exercises, and for years fourteen, sixteen, and eighteen there are six regular exercises with alternates for each year.² The type of performance varies considerably with the different exercises, hence it is best described by quoting certain typical ones. These exercises are presented to the child by means of spoken directions. The examiner tests one child at a time. The test should be given in a quiet room where there is freedom from distraction, and a friendly attitude between examiner and subject should be maintained. The examiner is directed to make sure that the subject understands what is to be done, and in all cases the burden of proof is with the examiner to

54 tests, we have eliminated 3 and relocated 32, leaving only 19 in positions assigned by Binet. The 3 eliminated are: repeating 2 digits, resisting suggestion, and 'reversed triangle.'

"The revision is really more extensive than the above figures would suggest, since minor changes have been made in the scoring of a great many tests in order to make them fit better the locations assigned them. Throughout the scale the procedure and scoring have been worked over and made more definite with the idea of promoting uniformity. This phase of the revision is perhaps more important than the mere relocation of tests."

¹ Terman, L. M., *The Measurement of Intelligence*. (Boston: Houghton Mifflin Company, 1916.)

Terman, L. M., and others, *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence*. (Baltimore: Warwick and York, 1917.)

² There are no tests for years eleven, thirteen, fifteen, and seventeen.

show that the subject has responded in a way that is representative of his ability. When the examiner has arranged the necessary materials,¹ he proceeds by giving the set of exercises for the highest year in which he judges the subject will pass all of them. If the child does these correctly, he passes to those for the higher years until the exercises become so difficult that the child is unable to do them. If the child fails to do all of the exercises of the first set, the examiner tries those for lower years until a set is found which is sufficiently easy for the child.

The following are illustrative of the sets of exercises used:

YEAR III

1. Points to parts of body. "Show me your nose, eyes, mouth, hair." Passed if 3 of 4 are correct. (Credit 2 months.)
2. Names familiar objects. "What is this?" Show in turn a key, a penny, a closed knife, a watch, a pencil. Passes if 3 of 5 are correct. (Credit 2 months.)
3. Pictures, enumeration or better. At least three objects in one picture. "Tell me everything you see in the picture." The three pictures used are, The Dutch Home, The Canoe, and The Post Office Scene. (Credit 2 months.)
4. Gives sex. If S* is a boy E asks, "Are you a little boy or a little girl?" If S is a girl, E asks, "Are you a little girl or a little boy?" (Credit 2 months.)
5. Give last name. "What is your last name?" (Credit 2 months.)
6. Repeats 6-7 syllables. Passed if 1 of 3 is correct.
 "Say just what I say."
 a. I have a little dog.
 b. The dog runs after the cat.
 c. In summer the sun is hot.
 (Credit 2 months.)

Alternate. Repeats three digits. 1 of 3. Order correct.
 Read 1 per second.

6-4-1 3-5-2 8-3-7

* S=Subject. E=Examiner.

¹ These are supplied by Houghton Mifflin Company.

YEAR IX

1. S must name (a) Day of week, (b) month, (c) day of month, and (d) year. Credit 2 months if a, b, and d are correct and error in c is not more than 3 days.
2. E presents five weighted cubes which all look alike. They weigh 3, 6, 9, 12 and 15 grams. S must arrange them in order of weight twice out of three trials. (Credit 2 months.)
3. S makes change without use of coins, paper or pencil. S must take 4¢ from 10¢; 12¢ from 15¢, and 4¢ from 25¢. Passed if 2 of 3 are correct. (Credit 2 months.)
4. S must repeat 4 digits backwards. E reads one per second.
6-5-2-8.....4-9-3-7.....8-6-2-9
Passed if 1 of 3 is correct. (Credit 2 months.)
5. S must make one sentence or not over two coördinate clauses which include the three words: (a) boy, river, ball; or (b) work, money, men; or (c) desert, rivers, lakes. Passed if 2 of 3 are correct. (Credit 2 months.)
6. Rhymes. S must give three rhymes for each word. One minute is allowed for each word. The words are: day, mill, and spring. Passed if three rhymes are furnished for each of two of the words. (Credit 2 months.)
Alternate 1. Name the months of the year. Time, 15 seconds.
Alternate 2. Give total value of stamps. Three 2 cent and three 1 cent stamps.

Scores secured. Two scores are secured by the use of the Stanford Revision of the Binet-Simon Scale. These are called "Mental Age" (M.A.) and "Intelligence Quotient" (I.Q.). The mental age is found by adding to the highest year in which all of the tests are passed the scores made in the following years. For example, a child passes all of the exercises in Year VIII, three in Year IX, two in Year X, and one in Year XII — then his score would be recorded as follows:

VIII.....	8 years 0 month
IX.....	—— 6 months
X.....	—— 4 months
XI.....	—— 2 months
Mental Age...9 years.	

If the child whose score we have here is ten years of age, the intelligence quotient would in this case be 0.9×100 , or 90. The formula for the I.Q. being $\frac{\text{mental age}}{\text{chronological age}} \times 100 = \text{I.Q.}$ The mental age score is interpreted to be the measure of the subject's general mental ability. Within the limits of the test's capacity to measure, all pupils of the same mental age are assumed to have the same ability to do ordinary school work, handle abstract relations, etc., regardless of their chronological ages. The intelligence quotient is interpreted to mean the ratio of the subject's mental age to the mental age of a typical or normal child of the same chronological age. If the subject's I.Q. is 100, this means that he has the mental age which is the average mental age of all of the children of his chronological age. In so far as this intelligence quotient remains constant, it is a prediction of the probable limits of development of the child's intelligence.

Advantages and limitations of the Stanford-Binet Scale. The chief advantage of an individual mental examination is the opportunity it affords for a trained examiner to observe the child's behavior under standardized conditions. When giving group tests the difficulties and mistakes of individuals are not so readily observed. In a group test the subject often is led to guess or waste time on misunderstood or very difficult exercises. In the individual tests the examiner may guide the subject into the best possible response for that subject.

Along with this advantage there is the disadvantage that individual tests demand the service of more highly trained examiners. The minimum of training is probably that secured in the supervised testing of at least twenty children.¹

¹ For a discussion of the training of teachers for intelligence testing see Dickson, Virgil E., and Martens, Elise H., "Training Teachers for Mental Testing in Oakland, California"; in *Journal of Educational Research*, vol. 7, pp. 100-08. (February, 1923.)

Terman¹ describes the requirements in terms of personality and in training in addition to that secured from giving the tests. In general these personal qualities are those which are found in the very best teachers. Psychological training beyond that given in the standard teacher training courses is desirable, but not necessary.²

An examiner possessing the minimum qualifications should secure fairly reliable mental-age scores with several of the individual tests. Such examiners should always work under the supervision of trained psychologists. If working only under the direction of the supervisory staff of a school system, the results of the mental tests should never be offered as evidence for a diagnosis of any grade of feeble-mindedness. In all situations where low I.Q.'s are secured, the best attitude is that of doubt of the reliability of the score. The next step should be that of inquiry and investigation into possible causes for so low a score. Terman³ states that "Nearly all who are below 70 or 75 I.Q. are feeble-minded." Since the army testing experience there has been a strong tendency to modify this statement, as many soldiers with I.Q.'s of much less than 70 were found to have been self-

¹ Terman, L. M., *The Measurement of Intelligence*, pp. 124-27. (Boston: Houghton Mifflin Company, 1916.)

² For the person who cannot take formal training in the giving of Binet tests, these exercises will be valuable after a careful study of the manual. Observe a well-trained examiner. If possible observe such an examiner during the giving of tests which cover the whole field of the scale. This usually means the examination of three subjects, one of about three years of age, another of five or six years, and a third of about nine years of age. After such observations the novice should try giving the scale throughout with the manual at hand, preferably to an adult. After giving the scale two or three times to adults, the examiner should strive to correct errors in procedure which have been discovered, and then test several subjects beginning with younger children. Obviously these first tests should be regarded as practice tests and the results should not be considered as more than rough indications of the true measures.

³ Terman, L. M., *The Measurement of Intelligence*, p. 141. (Boston: Houghton Mifflin Company, 1916.)

supporting citizens and efficient soldiers. The diagnosis of feeble-mindedness, unless the case is very obvious, should be left to the psychiatrist or the psychologist with institutional experience.

The individual examination requires rather expensive material and a large expenditure of time for giving the examinations. Careful tests require from thirty minutes to an hour and a half, younger children usually requiring less time than older children and adults.

Reliability and validity of the Stanford-Binet Scale. The Stanford-Binet Scale has been generally assumed to give a highly valid and reliable measurement of intelligence. The reliability and validity of measures of intelligence by means of other instruments, particularly group tests, are frequently given in terms of a comparison with the results of Stanford-Binet examinations. This opinion is based partly upon the obviously careful way in which Binet scores are secured. The examiner is supposedly well trained, and is able to give full attention to one subject at a time. The thorough and careful standardization of the Stanford-Binet has added greatly to the feeling of confidence which examiners have in this scale. More tangible evidences of the degree of reliability have been furnished by a great number of students who have reported the results of the use of the scales. Some of these results are confirmatory of Terman's original study. For example, Miss Edith Whitcomb¹ shows that 2360 primary children, measured with the Stanford-Binet, show a distribution of I.Q.'s in striking agreement with Terman's distribution of 905 cases.

Other studies have been concerned with the relationship between Stanford-Binet scores and success in school work.

¹ Whitcomb, M. Edith, "Intelligence Tests in the Primary Grades"; in *Journal of Educational Research*, vol. 5, pp. 58-60. (January, 1922.)

Terman¹ reports a correlation of .45 between I.Q. and quality of school work. Miss Whitcomb found "that approximately eighty-five or ninety per cent do work that agrees fairly well with the intelligence quotient."² Terman reports a correlation of .48 between teachers' ratings of intelligence and I.Q. Neither of these relationships is conclusive, as both the school marks and the teachers' estimation of intelligence are inaccurate. A correlation of .60 is about all that we can expect between such criteria and a perfect measure of intelligence.

When the reliability of the Stanford-Binet Scale is studied by comparing the scores from half of the exercises of each age group with the scores on the other half, the resulting coefficient of correlation is above .90.³ This shows that the scale measures very consistently the ability or capacity which it measures.

The Stanford-Binet Scale does not measure normal or superior older children and adults as well as it measures younger children. This relatively inaccurate measurement of older subjects is due to the relatively low upper limit of the scale. A subject who passes all of the tests in the fourteen-year group has only the sixteen- and eighteen-year groups in which to gain additional credits, while a child who passes all the tests in the eight-year group, but fails in some tests in the nine-year group, still has a chance to make scattering credits throughout all of the higher groups.

Retests and constancy of the I.Q. Retests of the same children after an interval have frequently been used as an indication of the reliability of measurements by the Binet

¹ Terman, L. M., *The Stanford Revision and Extension of Binet-Simon Scale for Measuring Intelligence*, p. 127. (Baltimore: Warwick and York, 1917.)

² Whitcomb, M. Edith, "Intelligence Tests in the Primary Grades"; in *Journal of Educational Research*, vol. 5, p. 59. (January, 1922.)

³ The coefficient of reliability is slightly larger, about .95.

Scale. The I.Q.'s secured by the first and later examinations were compared and coefficients of correlation calculated. Baldwin and Stecher¹ summarize their studies as follows: "The coefficients of correlation between all examinations within the four groups are high and reliable, ranging from $+.72 \text{ P.E.} \pm .05$ to $+.93 \text{ P.E.} \pm .02$, showing that they may be used as a basis for prediction. The correlations are probably only slightly modified by the personal equations of the examiners." "For comparison it is of interest to note the size of the correlations obtained by other examiners. These were: Bobertag, $+.95$ (Binet), Terman $+.93$ (Stanford), Cuneo and Terman $.95$, $.94$ and $.85$ (Stanford), Rosenow $.82$ (Binet and Stanford), Rugg and Colloton $.84$ (Stanford)." These results are similar to those of certain unpublished studies which have come to the writer's attention. Since Baldwin and Stecher's coefficients were not corrected for attenuation, it seems reasonable to suppose that the true reliability is expressed by a coefficient of approximately $.93$. The results of these studies may be summarized by the statement that the chances are even that a second test will not change a child's I.Q. more than 4.5 points.

There is evidence that measurements by a trained examiner are highly objective. Terman² reports correlation of $.929$ and $.94$ between tests on the same children given by different examiners after an interval. Since these are fully as high correlations as are found on retests by the same examiners, it would indicate that the personal factor is not often operative in changing the results of the tests. Other workers agree with these conclusions.

¹ Baldwin, Bird T., and Stecher, L. I., *Mental Growth Curve of Normal and Superior Children*. University of Iowa Studies, vol. 2, no. 1, pp. 46 and 53. (Iowa City: University of Iowa, 1922.)

² Terman, L. M., *The Intelligence of School Children*, pp. 142-46. (Boston: Houghton Mifflin Company, 1919.)

Although the evidence indicates the reliability of the Stanford-Binet is high, examiners must be cautioned to be watchful for those exceptional cases in which one testing does not give a true measure of the child's intelligence. Baldwin and Stecher¹ report a case in which the I.Q.'s for successive tests were 111, 116, 139, 140, 138. They report that "a careful study of this case showed no difference in the method of examination and no unusual physical condition aside from the adolescent physiological acceleration." It is significant, however, that the first and second examinations were made by E. V., while two other examiners made the other tests. The writer has found cases in which one examiner could not secure a complete response from a subject from which other examiners secured excellent responses.

The investigation of the mental growth of children and the study of the results of retests has led to some controversy over the constancy of the I.Q.² Without attempting to review the controversial literature, we may state that, for practical purposes, the I.Q. may be assumed to be approximately constant for normal children within the age limits of three to twelve. Subnormal children and gifted children do not show as high a degree of constancy in the I.Q. As Kuhlman³ has pointed out, absolute constancy is not neces-

¹ Baldwin, Bird T., and Stecher, L. I., *Mental Growth Curve of Normal and Superior Children*. University of Iowa Studies, vol. 2, no. 1, p. 36. (Iowa City: University of Iowa, 1922.)

² Doll, Edgar A., "The Growth of Intelligence"; in *Journal of Educational Psychology*, vol. 10, pp. 524-25. (December, 1919.)

Terman, L. M., "Mental Growth and the I.Q."; in *Journal of Educational Psychology*, vol. 11, pp. 325-41. (June, 1921.)

Freeman, F. N., "The Interpretation and Application of the Intelligence Quotient"; in *Journal of Educational Psychology*, vol. 12, pp. 3-13. (January, 1921.)

Rugg, H., and Colloton, C., "Constancy of the Stanford-Binet I.Q. as Shown by Retests"; in *Journal of Educational Psychology*, vol. 12, pp. 315-22. (June, 1921.)

³ Kuhlman, F., "The Results of Repeated Mental Reexaminations of

sary for purposes of prediction. If we know the I.Q. is approximately constant for a certain period for normal children, and that it varies in known ways for other children and for other life periods, this is sufficient for purposes of prediction. The evidence which is accumulating points to this kind of behavior of the I.Q. The conclusion of Baldwin and Stecher¹ is generally accepted: "An analysis of the individual growth curve shows that the I.Q. is only approximately constant during successive examinations. The amount of difference between I.Q.'s obtained at various examinations is sufficiently small, and the correlations between the examinations are sufficiently high, with small probable errors of estimate, to permit of predicting from an earlier examination what the individual's later development will be."

Goddard's Revision of the Binet Scale (1911). This revision was the first one made in America, but it has been very largely displaced by the Stanford Revision. Goddard did not change the content of Binet's 1911 Scale, but shifted the tests about and added tests for Year XV and "Adult."² The testing procedure is similar to that of the Stanford Revision. Terman³ reports that scores secured from Goddard's Revision were higher than those secured by the Stanford Revision in the lower years, and lower in the years above ten or eleven. Making this allowance, the

639 Feeble-Minded over a Period of Ten Years"; in *Journal of Applied Psychology*, vol. 5, pp. 195-224. (September, 1921.)

¹ Baldwin, Bird T., and Stecher, L. I., *Mental Growth Curve of Normal and Superior Children*. University of Iowa Studies, vol. 2, no. 1, p. 59. (Iowa City: University of Iowa, 1922.)

² Goddard, H. H., "Two Thousand Normal Children Measured by the Binet Measuring Scale of Intelligence"; in *The Pedagogical Seminary*, vol. 28, pp. 232-59. (June, 1921.)

Goddard, H. H., *The Binet-Simon Measuring Scale for Intelligence*. Reprinted from the *Training School Manual*, January, 1920. (This Manual may be secured from the Training School, Vineland, New Jersey.)

³ Terman, L. M., *The Measurement of Intelligence*, p. 62. (Boston: Houghton Mifflin Company, 1916.)

reliability and validity of Goddard's Revision approximates that of the Stanford-Binet.

Kuhlman's Revision.¹ Kuhlman's early revision had the same faults as the Goddard Revision; that is, the scores were too high for the lower ages and too low at the upper end. Also like Goddard's and the original Binet, the directions were not explicit and standardized. Both of these defects are remedied in the 1922 Revision. Furthermore, Kuhlman has eliminated nineteen of the original tests which were unsatisfactory, and he has increased the number of tests in each age group to eight. This should add to the reliability of the scale. The scale was standardized by the examination of about seven thousand children. In the manual there are full explicit directions for giving the tests, and a table from which I.Q.'s may be read directly is given in the Appendix. Examiners trained in the use of the Stanford-Binet will need considerable practice before they are proficient in the use of the Kuhlman Revision, and *vice versa*.

An interesting variation of the Kuhlman Revision is the addition of tests for ages below the third year. All other revisions of the Binet Scale begin at three years. Kuhlman supplies tests for ages three months, six months, twelve months, eighteen months, and two years. Samples of these tests are given with only suggestive descriptions.

¹ Kuhlman, F., *A Revision of the Binet-Simon System for Measuring the Intelligence of Children*. (Monograph Supplement of *Journal of Psycho-Asthenics*, September, 1912. 41 pp.)

Kuhlman, F., "Some Results of Examining 1000 Public School Children, with a Revision of the Binet-Simon Tests of Intelligence by Untrained Teachers"; in *Journal of Psycho-Asthenics*, vol. 28, pp. 155-79, 233-69. (1914.)

Kuhlman, F., *A Handbook of Mental Tests*. (Baltimore: Warwick and York, 1922. 208 pp.)

FOR THE THREE-MONTHS-OLD CHILD ¹

1. Carrying hand or object to mouth. Place a small block or other object in the child's right hand and note if it is carried to the mouth. Repeat for left hand.
2. Reaction to sudden sounds.
3. Binocular coördination.
4. Turning eyes to object in marginal field of vision.
5. Winking at an object threatening the eyes.

AGE, TWELVE MONTHS

1. Sitting and standing.
2. Speech. Successful combination of two or three syllables.
3. Imitation of movements.
4. Marking with pencil.
5. Recognition of objects.

AGE, EIGHTEEN MONTHS

1. Drinking.
2. Feeding with spoon or fork.
3. Speech. Use of some words, or understanding of a question unaccompanied by gesture.
4. Spitting out solids.
5. Recognition of objects in pictures.

Advantages and limitations of the Kuhlman Revision. Kuhlman² reports retest results for a group of feeble-minded children which are similar to those secured by use of the

¹ Kuhlman, F., *A Handbook of Mental Tests*, pp. 86 ff. (Baltimore: Warwick and York, 1922.)

Examples quoted are inadequately described for testing purposes. Obviously one who would give these tests must follow the complete descriptions given in the Manual.

² Kuhlman, F., "The Results of Repeated Mental Reëxaminations of 639 Feeble-Minded over a Period of Ten Years"; in *Journal of Applied Psychology*, vol. 5, pp. 195-224. (September, 1921.)

Stanford Revision. Since the Kuhlman Scale is very similar to the Stanford-Binet in both content and in the directions for giving and scoring, we may estimate that the Kuhlman Revision will secure results similar in validity and reliability to those secured by the Stanford Revision.

The Kuhlman Scale may be used down to the three-months-old level, which makes it valuable for the testing of infants. If the standards here prove valid and reliable, this test should be of great service for those who wish to adopt foundlings or orphans. The Kuhlman Revision does not secure a mental age score above fifteen years. This restricts its usefulness for the older children.

2. *Individual performance tests*

The Binet Scales have often been criticized for making undue use of language. For meeting the criticism there have been a number of performance tests designed to test special functions. Such tests as the tapping tests, form boards, cube construction, and many others have long been in use in psychological clinics.¹

Pintner-Paterson performance tests.² The Pintner-Paterson Performance Scale is made up of a well-selected group of tests which involve a minimum of language. This scale was planned for use with deaf children, but has since been used successfully with foreign children and other children who have language difficulties. There are fifteen tests in all, most of which require the subject to fit blocks or parts of a picture into holes cut from a board or picture. This description, however, gives a very inadequate idea of their nature and use.

¹ See Whipple, G. M., *Manual of Mental and Physical Tests*, vols. I and II. (Baltimore: Warwick and York, 1910.)

² Pintner, Rudolf, and Paterson, Donald, *A Scale of Performance Tests*. (Baltimore: Warwick and York, 1917.)

The Army Performance Scale.¹ After considerable experimentation with various performance tests, the army examiners adopted ten groups of tests as a Performance Scale. These tests are named in Table XXVI, which gives the correlations between the tests of the Army Performance Scale² and the Stanford-Binet. The tests are arranged in the order of correlation rank; the numbers at the left denote the numbers of the tests in the Army Performance Scale.

TABLE XXVI. COEFFICIENT OF CORRELATION (r) BETWEEN STANFORD-BINET AND THE TESTS OF THE ARMY PERFORMANCE SCALE. (P.P. = PINTNER-PATERSON SCALE.)

NO. OF TEST ON ARMY SCALE	NAME OF TEST	r	NO. OF CASES
7	Digit symbol	0.777	260
6	Designs	0.735	260
9	Picture arrangement	0.723	134
2	Mannikin and feature profile (P.P.)	0.676	260
1	Ship test (P.P.)	0.661	134
8	Maze test (Porteus)	0.655	260
10	Picture completion	0.650	134
4	Cube construction	0.633	260
3	Cube imitation (P.P.)	0.597	134
5	Form board	0.480	134

Scores from the complete Performance Scale correlated .834 with scores from the Stanford-Binet. As the army examiners had tried the Porteus Mazes, the Pintner-Paterson Performance Scale, and various other combinations of non-language tests, it is probable that the Army Performance Scale represents an improvement over the other scales.

¹ Yerkes, Robert M., *Memoirs of the National Academy of Sciences*, vol. 15, part 2, pp. 400-07. (Washington: Government Printing Office, 1921.)

² Adapted from *Memoirs of the National Academy of Sciences*, vol. 15, part 2, p. 404. (Washington: Government Printing Office, 1921.)

Such tests will find their chief usefulness in the hands of experienced clinicians.

3. Group tests for literate school children

Group tests for measuring intelligence include all of those which may be given by one examiner to several children or adults at the same time. This characteristic distinguishes them from the tests described in the preceding pages. In this section we shall describe those tests designed for literate school children who are able to read, with a reasonable degree of fluency, exercises printed in English and to write the responses asked for. In the following section we shall describe those group intelligence tests which do not require that a person taking the test be able to read.

Definition of terms. In describing group tests of intelligence, the word "examination" (or "scale") will be used to designate the booklet which usually includes several pages of tests and which is to be given at a sitting. The word "exercise" will be used for a single unit, such as, "How many are thirty men and seven men?" The word "test" will be used to denote a group of exercises usually on one page of an examination booklet, and usually worked out in response to one set of instructions without interruption. The word "form" will be used to denote a collection of material which constitutes an examination which is in a duplicate arrangement with another examination prepared by the same author or authors. The two examinations differ only in material used, and the scores of the two examinations are equivalent or capable of being corrected to an approximate equivalence. Thus "How many are 30 men and 7 men?" is the first "exercise" of the second "test" of group examination Alpha, Form 6. Obviously this use of terms cannot be followed where any of the terms are used with other meanings in titles of an examination.

Historical. The psychological examination of the men in the National Army gave a great impetus to the development and use of group examinations of intelligence. Prior to 1917 there were no group examinations of intelligence in general use. Doubtless many workers had in mind the preparation of intelligence tests which could be used with groups of children, and obviously the use of educational tests suggested this procedure. Some had used certain of the Binet tests with groups. Otis was working out a set of tests which would be roughly the equivalents of the best tests of the Stanford-Binet, and also be of such organization that they could be given to groups of children.

The following paragraph from the account of the construction of group examination used in the army indicates the influence of the pioneer work of Otis upon subsequent developments in this field:

Fortunately certain members of the committee ¹ had had encouraging experience with various types of group tests, and believed that these or others of similar kind could be readily adapted for army work. In this connection the contribution made by Arthur S. Otis, in devising a system of group tests, deserves special mention. The Otis tests embodied certain ingenious devices which permitted responses to be given without writing, and made possible objectivity in scoring. Otis generously placed all of his methods, together with correlation data which they had yielded, in the hands of Terman, who brought them before the committee. The scale which resulted from the committee's work bears a close resemblance to the Otis Scale. Four of the ten tests in the original Army Scale for Group Testing were taken from the Otis Scale practically without change, and certain others were shaped in part by suggestions derived from the Otis series.

¹ *Memoirs of the National Academy of Sciences*, vol. 15, p. 299. (Washington, 1921.)

The committee on the psychological examination of recruits was composed of R. M. Yerkes, Chairman, W. V. Bingham, Secretary, H. H. Goddard, T. H. Haines, L. M. Terman, F. L. Wells, and G. M. Whipple. It met at the Training School, Vineland, New Jersey, on the afternoon of May 28, 1917.

Army Alpha Group Intelligence Test. The Army Alpha Group Intelligence Test is the prototype of many of this group of examinations. For this reason, as well as for its historical interest, a more extended account of its derivation and structure will be given.

The committee in charge of the construction of the army group examination adopted twelve criteria for judging the appropriateness of tests which had been proposed for use. They were as follows:

(1) Adaptability for group use. (2) Validity as a measure of intelligence; that is, its correlation with other measures of intelligence of known validity. (3) The range of intelligence measured. It was extremely desirable to secure tests which would measure ability from the upper grades of mental deficiency to very superior levels of intelligence. This criterion made it impossible to consider certain tests which have shown themselves to be excellent measures of intelligence in somewhat restricted ranges. (4) Objectivity of scoring. It was agreed that, if possible, the tests should be arranged so that the responses could be scored by means of stencils. Otis at Stanford University and Thurston at the Carnegie Institute of Technology independently devised the stencil methods of scoring intelligence tests at approximately the same date (1915). Thorndike, however, had used it as early as 1914 in the scoring of a reading test. Otis seems to have been the first to arrange a battery of intelligence tests so that they could be scored exclusively by stencils. (5) Rapidity of scoring. (6) Unfavorableness to coaching. It was proposed to select only tests which could be made up in a large number of "forms" which would be entirely different in content but equal in difficulty. (7) Unfavorableness to malinering. (8) Unfavorableness to cheating. (9) Independence of schooling. It was agreed that the aim should be to test native ability rather than the results of school training. (10) Minimum of writing in responses. The aim was to secure tests which could be responded to in the main by underscoring, crossing out, etc., as in the majority of the Otis tests. (11) Interest and appeal. Everything else being equal the more interesting test should be preferred. (12) Economy of time.¹

¹ *Memoirs of the National Academy of Sciences*, vol. 15, pp. 299-300. (Washington: Government Printing Office, 1921.)

Ten tests were agreed upon by the committee with the criteria named in mind. The result of their work was Examination a, which is made up of ten tests as follows:

Test 1. Oral directions	Test 6. Synonym antonym
Test 2. Memory for digits	Test 7. Practical judgments
Test 3. Disarranged sentences	Test 8. Number of series completion
Test 4. Arithmetical reasoning	Test 9. Analogies
Test 5. Information	Test 10. Number comparison

After a thorough trial, Examination a was revised and the revision called Group Examination Alpha. In this revision the test of "memory for digits" (Test 2) and the test of "number comparison" (Test 10) were dropped from Examination a and the other eight tests somewhat changed. At about the same time the committee produced Examination Beta for illiterates. This examination was given by pantomime, and required no knowledge of the English language on the part of those examined. Both Alpha and Beta have been used to examine school children.

Army Group Intelligence Examination Alpha was found to possess the power to discriminate between the levels of intelligence of the adult men who were more intelligent and better readers, and the poorest twenty-five per cent in intelligence and literacy. It was too difficult for the lower one fourth in intelligence and literacy. It was sufficiently difficult so there were practically no perfect scores, but it was not very discriminating in its classification of the highest ten per cent of those examined. Since it was prepared for men, it is thought to be less adapted for women and girls. It is not so well adapted to the intermediate grades as are the National Intelligence Tests and other recently devised intelligence examinations.

The National Intelligence Tests. The most promising of the early revisions of the Army Group Intelligence Exam-

ination Alpha was the National Intelligence Tests. In describing their derivation Whipple¹ says:

Before the War, Professor Yerkes and Professor Terman approached the General Education Board for the support of a sort of school survey which would include the measurement of the intelligence of a good-sized group of pupils. The success of the Army Alpha Intelligence Examination made it evident that the same general methods would be applicable for such an examination of intelligence and that there would almost certainly be attempts made on the part of various individuals who had had contact with the army methods to adapt these to the examination of school children. It was felt that it would be very advantageous to the whole movement of mental testing if this adaptation could be made carefully, systematically, under the auspices of some institution or organization with prestige, and by men who would make a serious and expert contribution. The General Education Board acted favorably upon these suggestions with the proviso that the National Research Council should take the responsibility for the undertaking and that a group of four or five psychologists should coöperate in working out the details. A sum of money was appropriated for the work, and Messrs. Haggerty, Terman, Thorndike, Yerkes, and the speaker were made the members of the committee.

Types of tests in group intelligence examinations of the Army Alpha Type. As we have indicated, there are a number of intelligence examinations which are of the same general type as the Army Group Examination Alpha. In order to economize space, the general structure of the following group intelligence examinations will be described together: (1) Army Group Intelligence Examination Alpha; (2) Haggerty Intelligence Examination, Delta 2; (3) Illinois General Intelligence Scale; (4) National Intelligence Tests, Scale A and Scale B; (5) Terman Group Test of Mental Ability; and (6) Otis Group Intelligence Scale, Advanced Examination.

¹ Whipple, G. M., "The National Intelligence Tests"; in *Journal of Educational Research*, vol. 4, pp. 16-31. (June, 1921.)

The types of exercises which make up the tests of these examinations will be illustrated by quoting the fore exercises which precede the test proper or by giving representative exercises. Slight variations occur in the general structure of certain of the tests in different examinations. However, the description given here will serve to acquaint the student with the general character of the tests. For a more intimate acquaintance, one should secure a copy of the examinations and make a first-hand study of the tests. In all of the tests the exercises are arranged in approximately ascending order of difficulty, and the time allowed the pupils is such that practically none are able to finish. Sixteen types of tests are illustrated below.

1. *The Analogies Test.*

Look at this line:

(a) sky, blue:: grass — table green warm big

Notice the four words in heavy type. One of them — green — has a line drawn under it. Grass is green just as the sky is blue.

Look at the line (b) below: A fish swims and a man does what? Draw a line under the one word of the four in heavy type which tells what a man does.

Now look at line (c). Night means the opposite of day. What word means the opposite of white? Draw a line under it.

(b) fish — swims:: man — paper time walks girl

(c) day — night:: white — red black clear pure

2. *Arithmetic computation.* This is a general power test, similar to the Woody-McCall Arithmetic Scale (see page 56). There is, however, a time limit which is such that practically no student can complete all of the exercises.

3. *Arithmetic problems.* A test consisting of simple arithmetic problems is found in all of this group of intelligence examinations. The quantities in the problems are so small that practically no written computations are required.

4. *The Best Answer Test.*¹

Read each question or statement and make a cross before the BEST answer.

Sample

Why do we buy clocks? Because

1. We like to hear them strike.
2. They have hands.
3. They tell us the time.

Spokes of a wheel are often made of hickory because

1. Hickory is tough.
2. It cuts easily.
3. It takes paint nicely.

5. *Classification Test.*² In this test a pupil is given a list of five words, four of them are alike in some respect. The pupil is asked to find the word which is unlike the others and to cross it out. The following are sample exercises.

1. bullet cannon gun sword pencil
2. Canada Chicago China India France
3. Frank James John Sarah William

6. *Comparison Test.*

If the two things in a pair are the same, write S on the dotted line between them. If they are different, write D on the dotted line between them. Do each one as you come to it.

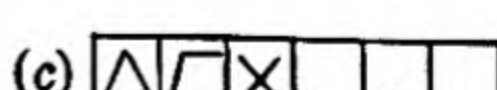
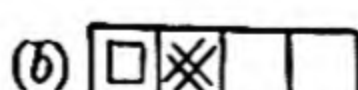
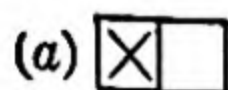
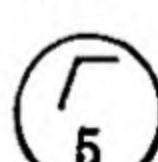
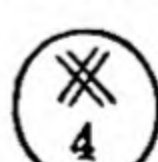
Begin here 273.....273

2861.....3854

Roland R.C.....Rollan R.C.

7. *The Digit Symbol Test.*³

Look at the sign and figure in each of the following circles:



¹ Also called Common Sense Test.

² This is very similar to the "similarity test" used by Otis.

³ Also called the Substitution Test.

Under the circles are some exercises having the same signs. Look at the exercise (a). Find the circle in which this sign is printed. The figure 3 is in the same circle. This means that the sign in exercise (a) stands for 3. Write the figure 3 in the square next to the sign to which it belongs. Look at exercise (b). There are two signs in this exercise. Find the figure which is in the same circle with the first sign. Write this figure in the first blank square. Do the same for the second sign. Look at exercise (c). Write the figures for these signs in the three blank squares. Write them in the order that the signs come.

8. *The Disarranged or Dissected Sentence Test.*¹

The words on each line below make one sentence if put in order. If the sentence the words would make is *true*, underline the word "true" at the side of the page. If the sentence they would make is *false*, underline the word "false."

men money for work.....(true false)

uphill rivers flow all.....(true false)

ocean waves the has.....(true false)

9. *Disarranged Sentence Test with Cross-out Variation.*²

Look at line (a) below. The words of this line are "see a I man on." In this order the words do not make sense but they can be made into a sentence if you leave out one word. The sentence is "I see a man." The word to be left out is "on." Draw a line through it. In each of the other lines when one word is crossed out, the remaining words can be made into a true sentence. Cross out the extra word in each line.

(a) see a I man on

(b) knife chair the sharp is

(c) John broken window trees has the

10. *Directions Test.* In the exercises of this test a child is given directions for doing something. In the more difficult exercises these directions are involved and intricate. The thing to be done is very simple. The following is a sample exercise.

¹ Also called the Mixed Sentence Test.

² Also called Verbal Ingenuity.

If 5 is more than 3, then cross out the number 4 unless four is more than 6, in which case draw a line under the number 5.

1 2 3 4 5 6 7 8 9

A certain letter is the fourth letter to the right of another letter. This other letter is midway between two other letters. One of these last two letters is next after E in the alphabet and the other is just before K in the alphabet. What is the "certain letter" first mentioned? (This exercise refers to a list of letters which is given as a part of the test.)

11. *The Information Test.*

Draw a line under the ONE word that makes the sentence true, as shown in the sample.

Our first President was

Adams Jefferson Lincoln Washington.

Coffee is a kind of

bark berry leaf root.

The logical selection test in the Terman Group Test and the National Intelligence Tests is a modification of this test which calls for the underlining of two words.

12. *Number Series Completion.*

5 10 15 20 25
20 18 16 14 12

In each row try to find out how the numbers are made up, then on the two dotted lines write the two numbers that should come next.

The Arithmetic Ingenuity Test in the Illinois General Intelligence Scale resembles this, but uses the cross-out modification.

13. *Picture Completion.*

Each of these pictures has something missing, and you are to put in with your pencil the missing part. Look at the first one. It is the picture of a boy's face, but it has no mouth. Now with your pencil mark in a mouth. The woman has no eye. Give her an eye. The other pictures are to be finished in the same way.



A



B

14. *Sentence Completion Test.*

Write on each dotted line one word to make the sentence sound sensible and right.

Sugar sweet.

Fire is hot, but ice is

15. *Sentence Meaning.*

Draw a line under the right answer.

Is coal obtained from mines?.....	Yes	No
Are all men six feet tall?.....	Yes	No
Does a conscientious person ever make mistakes?..	Yes	No
Are tentative decisions usually final?.....	Yes	No
Is rancor usually characterized by persistence?....	Yes	No

16. *The Synonym-Antonym Test.*

Look at these exercises:

- (a) good — bad same opposite
 (b) little — small same opposite
 (c) rich — poor same opposite

In exercise (a) "good" means the opposite of "bad." This is shown by a line drawn under the word "opposite." In exercise (b) "little" means the same as "small." Would you draw a line under "same" or "opposite"? You would draw it under "same." In exercise (c) do "rich" and "poor" mean the same or opposite? Draw a line under "same" or "opposite" to show your answer.

Table XXVII has been prepared to show the frequency of the occurrence of each type of test in the six group intelligence examinations listed on page 355. This table is to be read as follows: In the Army Alpha, Test 7 is "analogies"; Test 2, "arithmetic problems"; Test 3, "best answer"; Test 5, "disarranged sentences," etc. Slight modifications have been made in some of the tests by the authors of the examinations. From this table it will be noted that "arithmetic problems" occur in all except the National Intelligence Test, Scale B. "Arithmetic computation" occurs only in the examination just mentioned. Other tests which occur

frequently are "analogies," "information," and "synonym-antonym."

TABLE XXVII. SHOWING THE LOCATION OF THE VARIOUS TESTS IN SIX INTELLIGENCE EXAMINATIONS

TYPE OF TEST	ALPHA	HAGGERTY DELTA 2	ILLI- NOIS	NATIONAL A	NATIONAL B	OTIS	TER- MAN
1. Analogies.....	7	..	1	..	4	7	7
2. Arithmetic Com- putation.....	..	..	..	..	1	..	..
3. Arithmetic Prob- lems.....	2	2	2	1	..	5	5
4. Best Answer.....	3	5	..	..	..	..	2
5. Classification....	..	..	..	..	..	8	9
6. Comparison.....	..	..	..	..	5	..	..
7. Digit Symbol....	..	..	4	5	..	..	..
8. Disarranged Sen- tence.....	5	..	..	..	..	3	8
9. Verbal Ingenuity	..	..	5	..	..	..	..
10. Directions.....	1	..	..	..	..	1	..
11. Information.....	8	6	3	3	2	..	4
12. Number Series Completion....	6	..	..	..	..	..	10
13. Picture Comple- tion.....	..	3	..	..	..	..	..
14. Sentence Com- pletion.....	..	..	..	2	..	9	..
15. Sentence Mean- ing.....	..	1	..	..	3	..	6
16. Synonym-Anto- nym.....	4	4	7	4	..	2	3
Proverbs.....	..	..	..	..	..	4	..
Memory.....	..	..	..	..	..	10	..
Geometrical forms.....	..	..	..	..	..	6	..
Arithmetic In- genuity.....	..	..	6	..	..	..	..

Other group intelligence examinations for literate school children. There are a number of other group intelligence examinations for literate school children which have been used in a large number of schools. Among the most important of these are the Pressey Intermediate Classification Test; Pressey Senior Classification Test; Chicago Group Intelligence Test; Miller Mental Ability Test; and Otis Self-Administering Tests of Mental Ability, Intermediate Examination, and Higher Examination. Space does not permit a description of these examinations. Several of the tests in these examinations are similar to those described.

Limitations and advantages. Directions for scoring are given in the manuals accompanying the examinations. In most cases the scoring is to be done by means of stencils, which makes it highly objective and reduces the time required to a minimum.

The outstanding advantage of these measuring instruments in comparison with those which must be administered to children individually is the great economy of time resulting from testing children in groups. There is an added advantage in that less training and experience is required of the examiner.

Although these intelligence examinations have been presented as suitable for Grades III to XII inclusive, it must be borne in mind that their functions are not identical. Some are more appropriate for use in the high school than others. On the other hand, certain ones have been specifically intended for use in the elementary schools. The Army Group Examination Alpha was designed to measure the intelligence of adults. It has been used to secure satisfactory measurements of third-grade children, but it is not recommended for use below the seventh grade. Haggerty Intelligence Examination, Delta 2, is adapted to Grades III to IX. The Illinois General Intelligence Scale is especially recommended for Grades VI, VII, and VIII. The National Intelligence Tests are designed for Grades III to VIII. The Otis Group Intelligence Scale is best suited to high school and college. The Terman Group Test of Mental Ability is designed for use in junior and senior high schools.

Under the head of validity two questions arise: (1) Does the examination measure a pupil's capacity to learn or does it measure some other trait? (2) How accurately does it measure the thing which it measures or what is its reliability? The second of these questions is answered by calculating the coefficient of reliability. The first is answered in

part by comparing the scores yielded by the group intelligence test with certain criteria. The most frequently used criteria are measures secured by some revision of the Binet-Simon Tests, teachers' estimates, and composite scores.

The reliability of the Army Group Examination Alpha is approximately .85. The coefficient of correlation between the measures of the mental ages yielded by the test and the mental ages as determined by the Binet Scale is .81. When teachers are given careful instructions for estimating intelligence, correlations between test scores and teachers' estimates will be in the neighborhood of .60. The probable error of measurement of the Army Group Examination Alpha is about thirteen points. Otis gives the coefficient of reliability for his scale of .967 and the probable error of measurement 13.7 points or three and one half months of mental age.¹ Correlation between measures yielded by the Terman Group Scale and Stanford-Binet Scores is about .80. The reliability coefficient of .875 has been obtained for the former. The reliability of the Illinois General Intelligence Scale is .92. The coefficient of correlation between scores yielded by it and those yielded by the Otis Group Intelligence Test or the National Intelligence Tests is approximately .80.²

Stenquist has reported coefficients of correlation with a composite score obtained from the following examinations: National Intelligence Tests, Scale A, Scale B; Haggerty Group Intelligence Examination, Delta 2; Otis Group Intelligence Scale, Advanced Examination; Myers Mental Measure; and the Kelley-Trabue Completion Language

¹ Otis, A. S., "An Absolute Point Scale for the Group Measurement of Intelligence"; in *Journal of Educational Psychology*, vol. 9, pp. 333-48. (June, 1918.)

² Monroe, Walter S., *The Illinois Examination*. University of Illinois Bulletin, vol. 19, no. 9, Bureau of Educational Research Bulletin, no. 6, p. 58. (Urbana: University of Illinois, 1921.)

Test. The following coefficients of correlation with this criterion are reported: ¹

Delta 2.....	.808
National A.....	.801
National B.....	.788
Otis.....	.680

Franzen has made a similar study with a composite score of thirteen examinations.² The following correlations with this composite score are quoted from his study:

National A.....	.93
Terman.....	.92
Otis.....	.92
Delta 2.....	.91
Illinois.....	.90
National B.....	.90

Coefficients of reliability or coefficients expressing a comparison with criterion measures of intelligence do not furnish any index of the presence of constant errors. They relate only to variable errors of measurement. As we have indicated, the measurement of intelligence by such tests as we have considered here is based on the assumption that all of the children tested have had approximately the same opportunities for acquiring the achievements measured. Whenever this assumption is not approximately true, we may expect constant errors to be introduced into the scores. This will happen whenever the children are given instruction which functions as coaching on the test. Such instruction may be given unintentionally in the course of the regular work of the school. In one case the average increase in mental age for a group of 134 children during a period of

¹ Stenquist, J. L., "A Case for the Low I.Q."; in *Journal of Educational Research*, vol. 4, pp. 241-54. (November, 1921.)

² Haggerty, M. E., "Intelligence Examination, Delta 2"; in *Journal of Educational Psychology*, vol. 14, p. 273. (May, 1923.)

six months was slightly more than four years.¹ Whenever deliberate coaching occurs, greater constant errors may be introduced.

In using group intelligence tests it should be borne in mind that mental ages of above fourteen or sixteen years do not have the same meaning in the case of all the tests. This results in the intelligence quotient having a modified meaning. In order to recognize this difference, Otis has used "index of brightness" (I.B.) instead of the intelligence quotient.

4. Non-verbal intelligence tests

The intelligence examinations to be described under this heading require no reading knowledge of the English language by those being examined. A few of the tests are of such a character that the instructions can be given in pantomime, which makes it possible to test children unacquainted with the English language. Most of the tests described here are designed for use in the kindergarten and in the primary grades. A few of them have been designed to be used also in grades above the third.

The Army Group Intelligence Examination, Beta. This was the first of this type to be devised. It consisted of the following eight tests:

- Test 1. The Maze
- Test 2. Cube Analysis (counting tests)
- Test 3. X-O series (completion tests)
- Test 4. Digit Symbol
- Test 5. Number checking (number comparison)
- Test 6. Pictorial Completion
- Test 7. Geometrical Construction
- Test 8. Memory for Designs

¹ Monroe, Walter S., *The Illinois Examination*. University of Illinois Bulletin, vol. 19, no. 9, Bureau of Educational Research Bulletin, no. 6, p. 69. (Urbana: University of Illinois, 1921.)

The following non-verbal intelligence examinations have been widely used: Dearborn Group Test of Intelligence, Series I, Revised Edition (Grades I to III); Dearborn Group Test of Intelligence, Series II, Revised Edition (Grades IV to IX); Haggerty Intelligence Examination, Delta 1 (Grades I to III); Myers Mental Measure (Grades I to XII); Pintner Non-Verbal Mental Tests (Grades II to VIII); Pressey Primary Classification Test¹ (Grades I and II); Detroit First-Grade Intelligence Test.

The tests by Dearborn have been found to yield relatively valid measures of intelligence, but require considerable time to administer and to score. The tests by Haggerty and Pressey are convenient to give and also to score. This is one reason for their extensive use. The Myers Mental Measure has been criticized by some investigators,² although the author (Myers) has pointed out certain fallacies in their reasoning.³ The test is recommended for use in Grades I to XII inclusive. This makes convenient comparisons between pupils in different grades.

5. Group tests for college students

The Army Group Intelligence Examination Alpha has been used extensively with college students. The Otis Group Intelligence Test, Advanced Examination, has also been used for this purpose. In addition certain tests have been designed for high-school graduates and college students.

¹ This is a revision of the original Pressey Primary Scale which has been very widely used.

² Henmon, V. A. C., and Streitz, Ruth, "A Comparative Study of Four Group Scales for the Primary Grades"; in *Journal of Educational Research*, vol. 5, pp. 185-94. (March, 1922.)

Stenquist, J. L., "Unreliability of Individual Scores in Mental Measurements"; in *Journal of Educational Research*, vol. 4, p. 350. (December, 1921.)

³ Myers, G. C., "Some Fallacies in Testing Intelligence Tests"; in *Journal of Educational Research*, vol. 7, pp. 84-86. (January, 1923.)

Some of these have been used in determining the fitness of students applying for entrance to colleges and universities. The Thorndike Intelligence Examination for high-school graduates is one of this type. It requires about three and one half hours to administer it. The scores are reported to correlate about .50 with scholarship records. The Roback Mentality Tests were used at Simmons College. They show a similar correlation of scholarship. The Brown University Psychology Examinations are the result of several years of experimentation by Colvin.¹

For further information in regard to the use of intelligence examinations in colleges and universities, the reader is referred to Chapters VIII, IX, and X of the *Twenty-First Yearbook of the National Society for the Study of Education*.² In Chapter X of this *Yearbook*, Whipple reports on the use of intelligence examinations in twenty-nine colleges and universities.

III. USING MEASUREMENTS OF GENERAL INTELLIGENCE

The uses of measures of general intelligence in connection with the work of the school may be grouped under three heads: (1) promotion and classification of pupils; (2) vocational and educational guidance; (3) interpretations of measures of achievement. The first two of these uses will be considered in Chapter XIII. The third one is included in Chapter X, which deals with testing programs. Because of the importance of having measures of intelligence at hand in interpreting measures of achievement, the Illinois Group Intelligence Scale has been combined with tests for measuring achievement in silent reading and arithmetic to form the Illinois Examination (see page 382). A similar combination

¹ Colvin, S. S., "Psychological Tests at Brown University"; in *School and Society*, vol. 10, pp. 27-30. (July 5, 1919.)

² Public School Publishing Company, Bloomington, Illinois. 1922.

of an intelligence examination and achievement tests has been made by Pintner (see page 383).

In using intelligence examinations, it is necessary to bear in mind the limitations which have been mentioned on pages 362 to 365. These instruments are not perfect, but when used intelligently they will be very helpful to teachers and supervisors in becoming better acquainted with their children.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. Rate a class of pupils as you judge them to rank in intelligence regardless of chronological age.
2. Secure the chronological ages of the pupils in the class rated and re-rate them trying to allow for their ages.
3. Check your ratings by scores obtained from intelligence examinations. (If possible study these ratings by the correlation method.)
4. Test a class using two different intelligence examinations. Study the pupils whose scores show the greatest differences by (1) examining the test papers and (2) by questioning them orally with reference to their general information as well as the school subjects they have studied. (3) If possible give these pupils an individual test.
5. Find an opportunity to give an intelligence examination to the most stupid person and the intellectually brightest person of your acquaintance.
6. Test all members of a family and compare their I.Q.'s.
7. If records are available try retests on children tested a year ago, and study the results.
8. Keep a careful notation of your specific difficulties in studying, making change, reading, meeting situations, etc. Try to combine some of these into a test of intelligence.
9. Which individual examination would you use on foreign children who cannot speak English?
10. Which individual and which group examination shows the most satisfactory reliability?
11. Which test shows the highest reliability?
12. Which exercise described in this chapter, in your judgment, best measures intelligence?

SELECTED BIBLIOGRAPHY

For additional references consult the bibliographies by Bell, Boardman, and Whipple given in the list below.

Armentrout, W. D. "Classification of Junior High-School Pupils by the Otis Scale"; in *Education*, vol. 8, pp. 83-87. (October, 1922.)

- Baer, Joseph A. "Comparison of I.Q.'s on Successive Tests, Illinois Examination"; in *Journal of Educational Research*, vol. 7, p. 80. (January, 1923.)
- Bagley, William C. "Educational Determinism; or, Democracy and the I.Q."; in *Educational Administration and Supervision*, vol. 8, pp. 257-72. (May, 1922.)
- Bagley, William C. "Professor Terman's Determinism: A Rejoinder"; in *Journal of Educational Research*, vol. 6, pp. 371-85. (December, 1922.)
- Baldwin, Bird T., and Stecher, Lorle I. *Mental Growth Curve of Normal and Superior Children, Studied by Means of Consecutive Intelligence Examinations*. University of Iowa Studies in Child Welfare, vol. 2, no. 1. (Iowa City: University of Iowa, 1922. 61 pp.)
- Bell, J. Carleton. "Mental Tests and College Entrance"; in *Journal of Educational Psychology*, vol. 10, pp. 168-69. (March, 1919.)
- Bell, J. Carleton. "Group Tests of Intelligence: An Annotated List"; in *Journal of Educational Psychology*, vol. 12, pp. 103-08. (February, 1921.)
- Benson, C. E. "The Results of the Army Alpha Test in a Teacher-Training Institution"; in *Educational Administration and Supervision*, vol. 7, pp. 348-49. (September, 1921.)
- Bird, Grace E. "A Test of Some Standard Tests"; in *Journal of Educational Psychology*, vol. 11, pp. 275-83. (May, 1920.)
- Boardman, Helen. *Psychological Tests—A Bibliography*. Bureau of Educational Experiments, Bulletin No. 6. (New York, 1918.)
- Breed, Frederick S. "The Status of Intelligence Tests"; in *The School Review*, vol. 30, pp. 242-44. (April, 1922.)
- Bridges, James W. "The Correlation between College Grades and the Alpha Intelligence Tests"; in *Journal of Educational Psychology*, vol. 11, pp. 361-67. (October, 1920.)
- Bright, Ira J. "The Intelligence Examination for High-School Freshmen"; in *Journal of Educational Research*, vol. 4, pp. 44-55. (June, 1921.)
- Caldwell, Helen Hubbert. "Adult Tests of the Stanford Revision Applied to College Students"; in *Journal of Educational Psychology*, vol. 10, pp. 477-88. (December, 1919.)
- Chambers, George Gailey. "Intelligence Examinations and Admission to College"; in *Educational Review*, vol. 61, pp. 128-37. (February, 1921.)
- Chassell, Clara F. "The Results of the Thorndike Intelligence Examination in the Senior Class of the Horace Mann High School for Girls"; in *School and Society*, vol. 15, pp. 511-13. (May 6, 1922.)
- Clement, J. A., and Smythe, W. E. "Intelligence Tests and the Marks of Scholarship Men in College"; in *Educational Administration and Supervision*, vol. 7, pp. 510-16. (December, 1921.)
- Cobb, Margaret V. "The Limits set to Educational Achievement by Limited Intelligence"; in *Journal of Educational Psychology*, vol. 13, pp. 449-64, 546-55. (November, December, 1922.)
- Cobb, Margaret V. "One Element in the Probable Error of a Mental Age

370 EDUCATIONAL TESTS AND MEASUREMENTS

- Measurement"; in *Journal of Educational Psychology*, vol. 13, pp. 236-40. (April, 1922.)
- Cobb, Margaret V., and Toops, H. A. "Note on a Method of Studying Causes of Increase in Alpha Scores"; in *School and Society*, vol. 15, pp. 706-08. (June 24, 1922.)
- Cole, L. W. "Prevention of the Lockstep in Schools"; in *School and Society*, vol. 15, pp. 211-17. (February 25, 1922.)
- Colloton, Cecile, and Rugg, Harold. "Constancy of the Stanford-Binet I.Q. as Shown by Retests"; in *Journal of Educational Psychology*, vol. 12, p. 315. (September, 1921.)
- Colvin, S. S. "Nature and Measurement of General Intelligence"; in *Journal of Educational Psychology*, vol. 12, pp. 136-39. (March, 1921.)
- Colvin, S. S. "The Use of Intelligence Tests"; in *Educational Review*, vol. 62, pp. 134-48. (September, 1921.)
- Colvin, S. S. "Some Recent Results Obtained from the Otis Group Intelligence Scale"; in *Journal of Educational Research*, vol. 3, pp. 1-12. (January, 1921.)
- Colvin, S. S., and MacPhail, A. H. "The Value of Psychological Tests at Brown University"; in *School and Society*, vol. 16, pp. 113-22. (July 29, 1922.)
- Coxe, Warren W. "Norms for the Otis Group Intelligence Scale"; in *Journal of Educational Research*, vol. 3, pp. 313-14. (April, 1921.)
- Cunningham, K. S. "Binet and Porteus Tests Compared. Examination of One Hundred School Children"; in *Journal of Educational Psychology*, vol. 7, pp. 552-56. (November, 1916.)
- Dearborn, Walter F. "The Nature and Measurement of Intelligence"; in *Journal of Educational Psychology*, vol. 12, pp. 210-12. (April, 1921.)
- Dearborn, Walter F., and Lincoln, Edward A. "How the Dearborn Intelligence Examination Standards were Obtained"; in *Journal of Educational Psychology*, vol. 12, pp. 295-97. (May, 1922.)
- Dearborn, Walter F. "The Intelligence Quotients of Adults and Related Problems"; in *Journal of Educational Research*, vol. 6, pp. 307-25. (November, 1922.)
- Dearborn, Walter F., and Lincoln, Edward A. "Revising the Dearborn Intelligence Examinations"; in *Journal of Educational Psychology*, vol. 14, pp. 39-46. (January, 1923.)
- Dickson, Virgil E. "What First-Grade Children can do in School as Related to What is Shown by Mental Tests"; in *Journal of Educational Research*, vol. 2, pp. 475-80. (June, 1920.)
- Dickson, Virgil E., and Norton, John K. "The Otis Group Intelligence Scale Applied to the Elementary School Graduating Classes of Oakland, California"; in *Journal of Educational Research*, vol. 3, pp. 106-15. (February, 1921.)
- Dickson, Virgil E., and Martens, Elise H. "Training Teachers for Mental Testing in Oakland, California"; in *Journal of Educational Research*, vol. 7, pp. 100-08. (February, 1923.)

- Doll, Edgar A. "The Growth of Intelligence"; in *Journal of Educational Psychology*, vol. 10, pp. 524-25. (December, 1919.)
- Downey, June E. "The Constancy of the I.Q."; in *Journal of Delinquency*, vol. 3, pp. 122-31. (May, 1918.)
- Freeman, F. N. "The Interpretation and Application of the Intelligence Quotient"; in *Journal of Educational Psychology*, vol. 12, pp. 3-13. (January, 1921.)
- Freeman, F. N. "The Nature and Measurement of Intelligence"; in *Journal of Educational Psychology*, vol. 12, pp. 133-36. (March, 1921.)
- Freeman, F. N. "The Mental Age of Adults"; in *Journal of Educational Research*, vol. 6, pp. 441-44. (December, 1922.)
- Fryer, Douglas. "Occupational-Intelligence Standards"; in *School and Society*, vol. 16, pp. 273-77. (September 2, 1922.)
- Gambrill, Bessie Lee. "Some Administrative Uses of Intelligence Tests in the Normal School"; in *Twenty-First Yearbook of the National Society for the Study of Education*, Part II, pp. 223-43. (Bloomington, Illinois: Public School Publishing Company, 1922.)
- Garrison, S. C. "Fluctuation of Intelligence Quotient"; in *School and Society*, vol. 13, pp. 647-49. (June 4, 1921.)
- Garrison, S. C. "Additional Retests by Means of the Stanford Revision of the Binet-Simon Tests"; in *Journal of Educational Psychology*, vol. 12, pp. 307-12. (May, 1922.)
- Garrison, S. C., and Tippet, James S. "Comparison of the Binet-Simon and Otis Tests"; in *Journal of Educational Research*, vol. 6, pp. 42-48. (June, 1922.)
- Gordon, Kate. "Some Retests with the Stanford-Binet Scale"; in *Journal of Educational Psychology*, vol. 13, pp. 363-65. (September, 1922.)
- Guiler, Walter S. "How Different Mental Tests Agree in Rating Children"; in *Elementary School Journal*, vol. 22, pp. 734-44. (June, 1922.)
- Haggerty, M. E. "Tests of Applicants for Admission to University of Minnesota Medical School"; in *Journal of Educational Psychology*, vol. 9, pp. 278-86. (May, 1918.)
- Haggerty, M. E. "Recent Developments in Measuring Human Capacities"; in *Journal of Educational Research*, vol. 3, pp. 241-53. (April, 1921.)
- Haggerty, M. E. "Intelligence and its Measurements"; in *Journal of Educational Psychology*, vol. 12, pp. 212-16. (April, 1921.)
- Handschin, C. H. "Army Alpha in the Normal School"; in *School and Society*, vol. 13, pp. 476-77. (April 16, 1921.)
- Hansen, C. F., and Ream, M. J. "The Predicative Value of Short Intelligence Tests"; in *Journal of Applied Psychology*, vol. 5, pp. 184-86. (June, 1921.)
- Henmon, V. A. C. "The Nature and Measurement of Intelligence"; in *Journal of Educational Psychology*, vol. 12, pp. 195-98. (April, 1921.)
- Henmon, V. A. C. "The Measurement of Intelligence"; in *School and Society*, vol. 13, pp. 151-58. (February, 1921.)

- Herring, John P. "Verbal and Abstract Elements in Intelligence Examinations"; in *Journal of Educational Psychology*, vol. 12, pp. 511-17. (December, 1921.)
- Hicks, Vinnie Crandall. "The Value of the Binet Mental Age Tests for First-Grade Entrants"; in *Journal of Educational Psychology*, vol. 6, pp. 157-66. (March, 1915.)
- Johnston, J. B. "Tests for Ability before College Entrance"; in *School and Society*, vol. 15, pp. 245-53. (April 1, 1922.)
- Kuhlman, F. "Mentality Tests"; in *Journal of Educational Psychology*, vol. 7, pp. 280-82. (May, 1916.)
- Kuhlman, F. "The Results of Repeated Mental Reexaminations of 639 Feeble-Minded over a Period of Ten Years"; in *Journal of Applied Psychology*, vol. 5, pp. 196-224. (September, 1921.)
- Kuhlman, F. *A Handbook of Mental Tests; A Further Revision and Extension of the Binet-Simon Scale*. (Baltimore: Warwick and York, Inc., 1922. 208 pp.)
- Lincoln, Edward A. "The Constancy of Intelligence Quotients (a case study)"; in *Journal of Educational Psychology*, vol. 13, pp. 484-95. (November, 1922.)
- Lincoln, Edward A. "Time-Saving in the Stanford-Binet Test"; in *Journal of Educational Psychology*, vol. 13, pp. 94-97. (February, 1922.)
- Mitchell, David. "Psychological Examination and Pre-School Age Children"; in *School and Society*, vol. 15, pp. 561-68. (May 20, 1922.)
- Mohlman, Dora Keen. "The Discriminative Value of the Sub-Tests of a Group Intelligence Test"; in *School and Society*, vol. 15, pp. 399-400. (April 8, 1922.)
- Monroe, Walter S. "Some Correlations between Otis Scale and Rogers Mathematical Tests"; in *Journal of Educational Research*, vol. 2, pp. 774-76. (November, 1920.)
- Monroe, Walter S. *The Illinois Examination*. University of Illinois Bulletin, vol. 19, no. 9, Bureau of Educational Research Bulletin, no. 6. (Urbana: University of Illinois, 1921. 70 pp.)
- Moore, Henry T. "Three Types of Psychological Rating in Use with Freshmen at Dartmouth"; in *School and Society*, vol. 13, pp. 418-20. (April 2, 1921.)
- Murdoch, Katherine, and Sullivan, Louise R. "Some Evidence of an Adolescent Increase in the Rate of Mental Growth"; in *Journal of Educational Psychology*, vol. 13, pp. 350-56. (September, 1922.)
- Myers, Garry Cleveland. "Validating Intelligence Tests"; in *School and Society*, vol. 16, pp. 612-14. (November 25, 1922.)
- Myers, Garry Cleveland. "Some Fallacies in Testing Intelligence Tests"; in *Journal of Educational Research*, vol. 7, pp. 84-86. (January, 1923.)
- Odell, C. W. "Correlation of Certain Intelligence Tests for the Lower Grades"; in *Journal of Educational Research*, vol. 3, pp. 308-10. (April, 1921.)

- Otis, Arthur S. "An Absolute Point Scale for the Group Measurements of Intelligence, Part I"; in *Journal of Educational Psychology*, vol. 9, pp. 239-61, 333-48. (May, June, 1918.)
- Peterson, Joseph. "The Nature and Measurement of Intelligence"; in *Journal of Educational Psychology*, vol. 12, pp. 198-201. (April, 1921.)
- Peterson, Joseph. "The Growth of Intelligence and the Intelligence Quotient"; in *Journal of Educational Psychology*, vol. 12, pp. 148-54. (March, 1921.)
- Pintner, Rudolf, and Marshall, Helen. "A Combined Mental-Educational Survey"; in *Journal of Educational Psychology*, vol. 12, pp. 32-43. (January, 1921.)
- Pintner, Rudolf. "Nature and Measurement of Intelligence"; in *Journal of Educational Psychology*, vol. 12, pp. 139-43. (March, 1921.)
- Pintner, R. "The Significance of Intelligence Testing in the Elementary School"; in *Twenty-First Yearbook of the National Society for the Study of Education*, Part II, pp. 153-67. (Bloomington, Illinois: Public School Publishing Company, 1922.)
- Pintner, Rudolf, and Cunningham, Bess V. "The Problem of Group Intelligence Tests for Very Young Children"; in *Journal of Educational Psychology*, vol. 13, pp. 465-72. (November, 1922.)
- Porteus, S. D. "The Measurement of Intelligence: Six Hundred and Fifty-Three Children Examined by the Binet and Porteus Tests"; in *Journal of Educational Psychology*, vol. 9, pp. 13-31. (January, 1918.)
- Poull, Louise E. "Constancy of I.Q. in Mental Defectives, according to the Stanford-Revision of Binet Tests"; in *Journal of Educational Psychology*, vol. 12, pp. 323-24. (September, 1921.)
- Pressey, S. L. "Intelligence and its Measurement"; in *Journal of Educational Psychology*, vol. 12, pp. 144-47. (March, 1921.)
- Pressey, L. W. "A Group Scale of Intelligence for Use in the First Three Grades"; in *Journal of Educational Psychology*, vol. 10, pp. 297-308. (September, 1919.)
- Pressey, Luella W. "A Group Scale of Intelligence for Use in the First Three Grades: Its Validity and Reliability"; in *Journal of Educational Research*, vol. 1, pp. 285-94. (April, 1920.)
- Pyle, William Henry. *A Manual for the Mental and Physical Examination of School Children* (revised). University of Missouri Bulletin, vol. 21, no. 12. (Columbia: University of Missouri, 1920. 39 pp.)
- Richardson, Florence, and Robinson, Edward S. "Effects of Practice upon the Scores and Predictive Value of the Alpha Intelligence Examination"; in *Journal of Experimental Psychology*, vol. 4, pp. 300-17. (August, 1921.)
- Roberts, Alexander C. "Objective Measures of Intelligence in Relation to High-School and College Administration"; in *Educational Administration and Supervision*, vol. 8, pp. 530-40. (December, 1922.)
- Roberts, George L., and Brandenburg, G. C. "The Army Intelligence

374 EDUCATIONAL TESTS AND MEASUREMENTS

- Tests at Purdue University"; in *School and Society*, vol. 10, pp. 776-78. (December 27, 1919.)
- Rogers, Agnes L. "Intelligence Tests and Educational Progress"; in *Educational Review*, vol. 61, pp. 101-16. (February, 1921.)
- Rogers, Agnes L. "The Use of Psychological Tests in the Administration of Colleges of Liberal Arts for Women"; in *Twenty-First Yearbook of the National Society for the Study of Education*, Part II, pp. 245-52. (Bloomington, Illinois: Public School Publishing Company, 1922.)
- Root, W. T. "Two Cases Showing Marked Change in I.Q."; in *Journal of Applied Psychology*, vol. 5, pp. 156-58. (June, 1921.)
- Root, W. T. "The Intelligence Quotient from Two Viewpoints"; in *Journal of Applied Psychology*, vol. 6, pp. 267-75. (September, 1922.)
- Root, W. T. "Correlations between Binet Tests and Group Tests"; in *Journal of Educational Psychology*, vol. 12, pp. 286-92. (May, 1922.)
- Ruch, G. M., and Strachan, Lexie. "Intelligence Ratings by Group Scales and by the Stanford Revision of the Binet Tests"; in *Journal of Educational Psychology*, vol. 11, pp. 421-29. (November, 1920.)
- Rugg, Harold, and Colloton, Cecile. "Constancy of the Stanford-Binet I.Q. as Shown by Retests"; in *Journal of Educational Psychology*, vol. 12, pp. 315-22. (September, 1921.)
- Rusk, Robert R. "On Doctor Bagley's Rejoinder"; in *Journal of Educational Research*, vol. 7, pp. 269-70. (March, 1923.)
- Sandiford, Peter. "The Standardization of Tests and Scales"; in *Journal of Educational Research*, vol. 7, pp. 14-27. (January, 1923.)
- Scott, Walter Dill. "Intelligence Tests for Prospective Freshmen"; in *School and Society*, vol. 15, pp. 384-88. (April 8, 1922.)
- Smith, Franklin O. "The Relation between College Failures and Mental Tests"; in *School and Society*, vol. 16, pp. 444-46. (October 14, 1922.)
- Stenquist, John L. "Constancy of the Stanford-Binet I.Q. as Shown by Retests"; in *Journal of Educational Psychology*, vol. 12, pp. 54-56. (January, 1922.)
- Teagarten, Florence M. "'The Constancy of the I.Q.' again"; in *Journal of Educational Psychology*, vol. 13, pp. 366-72. (September, 1922.)
- Terman, L. M., Lyman, Grace, Ordahl, Dr. George, Ordahl, Dr. Louise, Galbreath, Neva, and Talbert, Wilford. "The Stanford Revision of the Binet-Simon Scale and Some Results from its Application to 1000 Non-Selected Children"; in *Journal of Educational Psychology*, vol. 6, pp. 551-67. (November, 1915.)
- Terman, L. M. *The Measurement of Intelligence*. (Boston: Houghton Mifflin Company, 1916.)
- Terman, L. M. "The Vocabulary Test as a Measure of Intelligence"; in *Journal of Educational Psychology*, vol. 8, pp. 452-66. (October, 1918.)
- Terman, L. M. "Some Data on the Binet Test of Naming Words"; in *Journal of Educational Psychology*, vol. 10, pp. 29-35. (January, 1919.)
- Terman, Lewis M. "The Use of Intelligence Tests in the Grading of School

- Children"; in *Journal of Educational Research*, vol. 1, pp. 20-32. (January, 1920.)
- Terman, L. M. "The Nature and Measurement of Intelligence"; in *Journal of Educational Psychology*, vol. 12, pp. 127-33. (March, 1921.)
- Terman, L. M. "Mental Growth and the I.Q."; in *Journal of Educational Psychology*, vol. 12, pp. 325-41, 401-07. (September, October, 1921.)
- Terman, L. M., and Whitmire, Ethel D. "Age and Grade Norms for the National Intelligence Tests, Scales A and B"; in *Journal of Educational Research*, vol. 3, pp. 124-32. (February, 1921.)
- Terman, Lewis M. "The Psychological Determinist; or Democracy and the I.Q."; in *Journal of Educational Research*, vol. 6, pp. 57-62. (June, 1922.)
- Theisen, W. W. "Does Intelligence Tell in First-Grade Reading?" in *Elementary School Journal*, vol. 22, pp. 530-34. (March, 1922.)
- Thorndike, Edward L. "Intelligence Examinations for College Entrance"; in *Journal of Educational Research*, vol. 3, pp. 329-37. (May, 1920.)
- Thorndike, Edward L. "The Reliability and Significance of Tests of Intelligence"; in *Journal of Educational Psychology*, vol. 11, pp. 284-87. (May, 1920.)
- Thorndike, Edward L. "On the New Plan of Admitting Students at Columbia University"; in *Journal of Educational Research*, vol. 4, pp. 95-101. (September, 1921.)
- Thurstone, L. L. "What is Meant by Intelligence"; in *Journal of Educational Psychology*, vol. 12, pp. 201-07. (April, 1921.)
- Thurstone, L. L. "Mental Tests for College Entrance"; in *Journal of Educational Psychology*, vol. 10, pp. 129-42. (March, 1919.)
- Trabue, M. R. "Some Pitfalls in the Administrative Use of Intelligence Tests"; in *Journal of Educational Research*, vol. 6, pp. 1-11. (June, 1922.)
- Tuttle, W. W. "The Thorndike Intelligence Examination as a Means of Determining Fitness for College"; in *Kentucky High School Quarterly*, vol. 8, pp. 14-28. (April, 1922.)
- Uhl, W. L. "Mentality Tests for College Freshmen"; in *Journal of Educational Psychology*, vol. 10, pp. 13-28. (January, 1919.)
- Wallin, J. E. Wallace. "The Results of Retests by Means of the Binet Scale"; in *Journal of Educational Psychology*, vol. 12, pp. 392-400. (October, 1921.)
- White, Wendell. "The Influence of Certain Exercises in Silent Reading on Scores in the Otis Group"; in *Educational Administration and Supervision*, vol. 9, pp. 179-82. (March, 1923.)
- Whipple, Guy M. "Mentality Tests"; in *Journal of Educational Psychology*, vol. 7, pp. 357-60. (June, 1916.)
- Whipple, Guy M. "The National Intelligence Tests"; in *Journal of Educational Research*, vol. 4, pp. 16-31. (June, 1921.)
- Whipple, Guy M. "Educational Determinism; A Discussion of Professor

Bagley's Address at Chicago"; in *School and Society*, vol. 15, pp. 599-602. (June 3, 1922.)

Whipple, Guy M. "Intelligence Tests in Colleges and Universities"; in *Twenty-First Yearbook of the National Society for the Study of Education*, Part II, pp. 253-70. (Bloomington, Illinois: Public School Publishing Company, 1922.)

Willard, Dudley W. "Native and Acquired Mental Ability as Measured by the Terman Group Test of Mental Ability"; in *School and Society*, vol. 16, pp. 750-56. (December 30, 1922.)

Woodrow, Herbert. "The Nature and Measurement of Intelligence"; in *Journal of Educational Psychology*, vol. 12, pp. 207-10. (April, 1921.)

Wylie, Andrew Tennant. "A Brief History of Mental Tests"; in *Teachers College Record*, vol. 23, pp. 19-33. (January, 1922.)

"The Correspondence between the Results of Illinois Intelligence and Binet-Simon Scales"; in *Journal of Educational Research*, vol. 6, pp. 274-75. (October, 1922.)

CHAPTER X

TESTING PROGRAMS

A testing program for a general survey of achievement. A single test of the type described in the preceding chapters measures a pupil's achievement in a relatively small part of the total field of school achievement. When one is desirous of securing a measure of the effectiveness of the school as a whole, or a pupil's average standing, it is necessary to give, not one standardized educational test, but a battery of them.¹ A testing program for such a purpose can be planned by selecting appropriate tests from those described in the preceding chapters, but certain difficulties will be encountered because the different tests have not been designed with special reference to their use together. In general each test has been constructed without reference to any other. In order to eliminate the necessity of selecting the tests for such a battery and the difficulties encountered in combining scores from several tests, certain groups of tests have been planned to be used together and have been published in a single booklet. There is an added convenience in having all of the directions for administering the tests in one manual and in being able to purchase all of the testing materials from one address.

A testing program should provide for comparison of achievement with general intelligence. When one wishes to inquire into the causes for a given degree of achievement, it is

¹ It should be remembered that we do not have standardized tests for measuring a number of important outcomes of instruction. (See Chapters VII and VIII.) Hence it is not possible to plan a testing program which will yield measures of *all* school achievements. Thus, the expression "effectiveness of the school as a whole" is used in a restricted sense.

necessary to consider the quality of the pupil material as well as the general organization of the school, the methods of instruction, and other factors which contribute to a child's achievement. A pupil who has little capacity to achieve cannot be expected to achieve as highly as those who have a much higher general intelligence. The same is true of classes. Occasionally we find assembled together in a class a group of pupils whose average general intelligence is unusually high. We also occasionally find classes of the opposite type. In interpreting measures of achievement of either groups of pupils or a single pupil, it is therefore necessary to obtain measures of their general intelligence.¹ At least two batteries of educational tests have included a test on general intelligence for this purpose.

Combining scores yielded by two or more tests. When making a general survey one wishes to secure a composite score which will serve as a single index of the total achievements of a pupil. Thus, when two or more tests are used one encounters the problem of combining the scores obtained. The magnitude of a score yielded by a standardized educational test depends upon the number of exercises, the time allowed for the test, and the plan of scoring. For this reason the scores yielded by different tests vary widely in magnitude. The rate score yielded by silent reading tests is frequently more than one hundred words per minute. The comprehension score yielded by the same test may be less than 15. In case of the Courtis Standard Research Test in Arithmetic, Series B, the maximum rate score is 24. Hence a score of any given magnitude, say 17, means one thing for one test and another thing for another. This condition is due to the lack of a common unit and a common zero point.

¹ For further discussion see Monroe, Walter S., *Introduction to the Theory of Educational Measurements*, pp. 245-49, 264-69. (Boston: Houghton Mifflin Company, 1923.)

Except by chance no two tests yield scores in terms of the same unit or expressed from the same zero point.

This makes it difficult to combine the scores yielded by different tests unless they happen to be constructed so that the units and zero points are approximately the same. In such a case the scores can be averaged without introducing a serious error. However, in general it is necessary to reduce the scores yielded by different tests to a common basis before they can be combined. A common method for reducing one set of scores to the scale of another set is to express each score as a deviation from the average in terms of the standard deviation as a unit.¹ This procedure is somewhat tedious and is not recommended except for investigations in which precision is highly desirable. It is possible to formulate a simple plan of weighting the scores yielded by the separate tests which will yield an average score sufficiently accurate for most purposes. When the zero points are approximately equivalent, a weighting inversely proportional to the ratio of corresponding norms may be used. For example, if the fifth-grade norm for Test A is 25 and that for Test B is 75, the scores yielded by Test A will be made approximately comparable to those yielded by Test B when the former are multiplied by 3. This plan of weighting assumes that equal weight is to be given to the two sets of scores.

A simple plan of weighting of this type has been used in the Monroe General Survey Scale in Arithmetic and in the Stanford Achievement Test. For practical purposes it is probably better to do this than to attempt to apply the more refined method. When the zero points of the tests are not approximately equivalent it will be necessary to add or subtract an appropriate amount from one set of scores.

¹ Monroe, Walter S., *Introduction to the Theory of Educational Measurements*. (Boston: Houghton Mifflin Company, 1923.) See pp. 211 ff. for a description of the method.

Derived scores. There have been several proposals for calculating a "derived score" which would be expressed in terms of the same unit for one test as for another and from the same zero point. Derived scores which satisfy this requirement may be combined by addition in exactly the same way as ordinary denominate numbers. Such scores are also more convenient to record and to interpret. The most widely used derived score is the "age score." The "point scores"¹ yielded by most intelligence tests are translated into mental ages. This is done by finding the average or median score which children of each age make. For example, if ten-year-old children make on the average a score of 72, then any child who makes a score of 72 is said to have a mental age of ten years. A similar plan has been followed in the case of a considerable number of achievement tests. The age norms furnish a basis for translating the point scores into achievement ages.

Various symbols have been used as abbreviations for achievement age in a particular subject. S.A. has been used to stand for "subject age." To indicate the subject age in a given subject the initial letter of the subject has been used in combination with "A." According to this plan A.A. would stand for "arithmetic age," R.A. for "reading age," etc. Thus, A.A. has been used for both "arithmetic age" and "achievement age." A pupil's average age in a group of subjects has been called his "educational age" (E.A.).

The quotient obtained by dividing the pupil's "educational age" (E.A.) by his "chronological age" (C.A.) has been called his "educational quotient" (E.Q.). For a given

¹ The term "point score" is used to refer to the score which is yielded directly by the test. The number of exercises done correctly, the number of exercises attempted, and the level of difficulty reached in the test are point scores.

subject we have a "subject quotient" (S.Q.), or, if it is wished to indicate a particular subject, R.Q. would be "reading quotient," A.Q. "arithmetic quotient," etc.

It has been proposed to go one step farther and divide a pupil's educational quotient or subject quotient by his intelligence quotient (I.Q.). This quotient has been called both "achievement quotient," and "accomplishment quotient" (A.Q.). A few test-makers have avoided using the word "quotient" for this purpose and called the result of the division the "achievement ratio" (A.R.). This relation can also be obtained by dividing the educational age or subject age by the mental age. The two procedures give identical results and are logically the same. In fact, the latter procedure is more frequently used.

In order to secure scores expressed in terms of a common unit and from a common zero point, McCall has proposed to translate point scores into T-scores.¹ The unit of the T-score is a measure of the variability of ability of twelve-year-old children. In order to obtain the basis for the translation of point scores into T-scores, it is necessary to have a test given to a large representative group of twelve-year-old children. Some of these will be found in the sixth grade and others in each of the two or three grades immediately below and immediately above the sixth. From the distribution of scores for these twelve-year-old children, it is easy to calculate the basis for making the translation. This method has considerable virtue in spite of certain limitations in the case of tests which are not suitable for twelve-year-old children. However, for practical purposes T-scores do not seem to be as useful as age scores.

¹ For an explanation of the T-score and its calculation see Monroe, Walter S., *Introduction to the Theory of Educational Measurements*, pp. 148 ff. (Boston: Houghton Mifflin Company, 1923.) See also McCall, W. A., *How to Measure in Education*. (New York: The Macmillan Company, 1922.)

The Illinois Examination. The Illinois Examination is a testing program consisting of the Illinois General Intelligence Scale, Monroe Standardized Silent Reading Test, Revised, and Monroe General Survey Scale in Arithmetic.¹ These three tests are printed in a single booklet of sixteen pages. For each of the achievement tests norms have been established for each half-year of mental age. These norms are used as a basis for translating the point scores into achievement ages. For example, the norm in arithmetic for a mental age of fifteen years is 68; for fifteen years and six months it is 72. When a pupil makes a score of 68 on the arithmetic scale he is said to have an achievement age in arithmetic of fifteen years. If his point score is 72, his achievement age is given as fifteen years and six months. The scores yielded by the general intelligence test are translated into mental ages. The quotient found by dividing a pupil's achievement age by his mental age ($\frac{A.A.}{M.A.}$) is called his "achievement quotient" (A.Q.). The achievement quotient is an index of the pupil's achievement in comparison with his capacity to achieve as measured by the intelligence test. Tables are provided for translating the point scores into age scores and also for calculating the achievement quotients.

The achievement quotients exhibit a somewhat greater degree of variability than intelligence quotients. Their distribution differs from the normal by showing a greater degree of variability above the median (approximately 100) than below the median. It is, however, possible to divide the distribution so that the per cent of pupils included in each division corresponds to that used for the interpretation of intelligence quotients. The following scheme for interpretation of quotients of individual pupils is suggested:

¹ See pages 355, 102 and 45 for descriptions of these tests.

QUALITY OF PUPILS' ACHIEVEMENT	ACHIEVEMENT QUOTIENT	PER CENT OF PUPILS INCLUDED
Very superior	{ 165 and above 135-164	1 6
Superior	117-134	13
Average	83-116	60
Poor	71- 82	13
Failure	{ 55- 70 Below 55	6 1

Pintner Educational Survey Tests. This group of achievement tests consists of eight separate tests. Five of them are abbreviations of the following tests: Thorndike Visual Vocabulary Scale, Woody Arithmetic Exercises, Kansas Silent Reading Tests, Thorndike Scale, Alpha 2, for the Understanding of Sentences, and Trabue Language Completion Tests. There are also tests on the following subjects: grammar, geography, and history. These are printed in a single booklet and the scores from the tests are to be added to form a measure of the pupil's average achievement. This battery of tests is designed to be used in connection with the Pintner Non-Language Mental Tests which are printed in a separate booklet.

Pintner has provided for translating the total point scores into a derived score called an "educational index." This is done by means of tables which are given in the Manual of Directions. This educational index is similar to McCall's T-Score, except that a different basis of translation is used for each age. McCall used the scores of twelve-year-old pupils for translating all point scores. Pintner uses a distribution of scores of eleven-year-old pupils for translating the scores made by eleven-year-old pupils, the distribution of scores made by ten-year-old pupils for translating the scores made by pupils of that age, and so on. He also provides for securing an educational index from the distribution

of scores by grades. The point score yielded by the non-language mental test is to be translated into a mental index by a similar procedure. The comparison between measures of achievement and measures of general intelligence is secured by subtracting the mental index from the educational index. This difference is somewhat comparable in significance to the achievement quotient yielded by the Illinois Examination and the accomplishment ratio yielded by the Stanford Achievement Test.

Stanford Achievement Test. The Primary Examination, designed for Grades II and III, consists of six tests: (1) Reading: paragraph meaning; (2) Reading: sentence meaning; (3) Reading: word meaning; (4) Arithmetic computation; (5) Arithmetic reasoning; and (6) Dictation exercise in spelling. The Advanced Examination, designed for Grades IV to VIII, consists of nine tests. The first five of the tests are similar to the first five tests of the Primary Examination. They, however, consist of more difficult exercises. The other four tests are as follows: (6) Nature study and science, (7) History and literature, (8) Language usage, and (9) Dictation exercise in spelling.

The paragraph reading test consists of a series of disconnected paragraphs from which one or more words have been omitted. It was intended to have the paragraphs of "such a nature that complete reading of the paragraph is necessary in order that the blanks may be correctly filled." The sentence reading test consists of a set of questions based upon general information which may be answered by "yes" or "no." The vocabulary test (word meaning) consists of exercises in which the pupil chooses one word out of a list of five to complete a sentence. The test on arithmetic computation is similar to the Woody-McCall Mixed Fundamentals Test. In the reasoning test, which is merely a list of arithmetic problems to be solved, an effort was made to

minimize the amount of computation necessary. Tests 6 and 7 of the Advanced Examination are intended to measure a pupil's information in nature study and science, and history and literature. Each of these two tests consists of ninety-five exercises in which the pupil is asked to answer a question by underlining one of three words or phrases. The following are samples:

- 36. A food rich in fats is....butter....eggs....tapioca....
- 37. An important meat-packing city is.....Chicago.....
New Orleans.....Seattle.....
- 42. A tree that will grow from cuttings is the....oak....pine
.....willow.....
- 64. The differential is a part of an.....auto.....bicycle
.....typewriter.....

In Test 8 the pupil is given sentences in which he is asked to mark the correct form. The following are samples:

- 10. I had ^{sat}_{set} there for an hour.
- 13. I think dominoes is an interesting ^{game.}_{sport.}
- 31. He acted the part ^{perfect.}_{perfectly.}
- 52. It is now ^{plain and evident}_{evident} why he left.

In all of the tests the exercises are arranged in ascending order of difficulty. The time allowances are sufficient for all pupils to try all of the exercises they "would have any considerable chance of answering." Hence the tests measure "power" only. The pupil's rate of work does not affect his score in any of the tests.

The exercises were selected with care. The principal criterion appears to have been that a satisfactory exercise must show an "increase from grade to grade in the per cent of pupils passing it. Items which did not show a marked increase over a range of at least three grades were elimi-

nated." In the construction of some of the tests the preliminary lists of exercises were based upon curriculum studies. This was true of spelling, vocabulary, and information (Tests 6 and 7). The reliability of the separate tests is high. For a composite score (average of scores on separate tests) the authors report a reliability coefficient of .98 for age groups.

A table has been prepared for translating the point scores into age scores. "Educational age" (E.A.) is the name given to the age equivalent of the composite score. The age equivalents of the scores yielded by the separate tests are called "subject ages," and for a particular test "arithmetic age" (A.A.), "reading age" (R.A.), etc., may be used. Educational age (E.A.) divided by chronological age (C.A.) is called "educational quotient" (E.Q.). In a similar manner subject quotients may be obtained. If the educational quotient (E.Q.) is divided by the intelligence quotient (I.Q.), the result is called the "accomplishment ratio" (A.R.). This is the same thing as the "achievement quotient" or "accomplishment quotient" (A.Q.).

Lippincott-Chapman Classroom Products Survey Tests. This group of tests consists of the following: (1) arithmetic fundamentals, (2) arithmetic problems, (3) reading selections test, and (4) reading continuous passage test. The two tests in the field of arithmetic have already been described. (See page 57.) The reading selections test is very similar to the Monroe Standardized Silent Reading Test, Revised. There are, however, a few exercises in which the pupil must write out his answer. The reading continuous passage test consists of a single passage to be read and twenty-seven questions based on it. Since thirty minutes are allowed for this one test, it is obvious that the pupil will have abundant opportunity to read and re-read the paragraph. All of the tests are designed to measure power

rather than rate. The point scores from the four tests are to be added to form the total score. This total score may in turn be translated into an educational age.

Pressey's Scale of Attainment No. 1. The Pressey Scale of Attainment No. 1 is designed to measure progress in reading, arithmetic, and spelling in the second grade. The spelling test consists of twenty-four words which are to be pronounced to the pupils who are required to write them on the test folder. The ability of pupils to recognize words is measured by means of exercises each of which consists of five groups of letters. Only one of these groups makes a word. The pupil is asked to find the word in each line and to draw a line around it. The test in arithmetic is based upon the fundamental combinations in addition and subtraction. Test IV for the understanding of sentences consists of exercises which form sentences when one word is omitted from each. For example, "I cat can see my doll." After the word "cat" is crossed out, the remaining words of this exercise form a sentence. The pupil is asked to cross out the extra word in each exercise.

Pressey Scale of Attainment No. 3. This battery of tests was designed to measure achievement in the third grade. The spelling test consists of sentences from which one word has been omitted. This word is given to the pupils and they are asked to write it in the blank space. The silent reading test consists of simple paragraphs to be read and simple questions about them to be answered. In this respect it is similar to Part II of the Courtis Silent Reading Test No. 2. (See page 106.) Four answers are given for each question. The pupil is asked to check the one which applies. The third part of the test is on arithmetic. Some exercises are questions on the fundamental combinations and some are simple problems to be solved. For each exercise four answers are given and the pupil is asked to check the one which is correct.

Pressey Scale of Attainment No. 2. This battery of tests is designed for measuring achievement in vocabulary, history, arithmetic, and English in the eighth grade. Test I — Reading vocabulary — consists of exercises of the following type:

A dungeon is a room in a:

(cathedral prison store museum.)

The student is asked to underline the word which makes the truest statement. Test II on American history consists of similar exercises. For example:

The people who settled in Plymouth were:

(Dutch English French German.)

Test III is a problem test in arithmetic. For each problem four answers are given. The student is asked to underline the one which is correct. Test IV, "good usage," consists of exercises such as the following:

The boys have ran away.

I've fell lots of times.

We was on time.

I rang the bell.

In each exercise one and only one sentence is correct. The student is asked to underline the one which is correct. Each of these four tests consists of forty exercises.

Selecting tests for a testing program. A well-defined purpose is a prerequisite to the formulation of a testing program. If one's purpose is to secure a detailed measurement of achievement in the field of a single subject, one should select tests with direct reference to that purpose. If one wishes to secure measures of both fluency and power, the testing program should include both a rate test and a power test. If one's purpose is to secure only a general measurement of the achievement within a given subject, it will probably be satisfactory to use a single test; as, for example,

the Woody-McCall Mixed Fundamentals Test in Arithmetic. No test should be included in a testing program unless it will contribute to the realization of one's purpose. As we have indicated on page 377, measures of general intelligence are necessary for certain types of interpretation. For this reason it will frequently be advisable to include a general intelligence test in the testing program. The batteries of tests described in the preceding pages have been assembled with certain purposes in mind. When one's purpose coincides with the implied purpose of one of these batteries of tests, it is advisable to use it rather than to plan a program involving the administration of another group of tests. Since each of these batteries of tests described in this chapter are bound in a single booklet, they will be found convenient to use. The directions are also printed in a single manual. There is a convenient record sheet for assembling the scores, and the scores from the separate tests can be easily combined to form a single general index of achievement. Furthermore, a battery of tests is accompanied by directions for interpreting the combined scores. This feature will usually be lacking in a specially planned testing program.

Cost of a testing program. There are few single tests which can be purchased for much less than one cent per pupil. The Illinois Examination sells for four cents per pupil. The Stanford Achievement Test is more expensive. However, one should bear in mind that the cost is for several tests rather than one. It is true that the cost of the test materials for a school system will appear large, but there is a corresponding "large" measure of achievement. The cost of testing material should always be considered in relation to the amount of information which is secured.

In addition to the outlay for testing materials, it is necessary to consider the time and cost of administering the tests and of scoring the test papers. Frequently teachers

complain concerning the time which they are required to devote to the scoring of test papers and voice the opinion that the information secured is not worth the investment which they have made. This has doubtless been true in some cases. In our endeavor to construct improved measuring instruments there has been much experimentation with tests which were crude and which were not planned with reference to ease and convenience of scoring. Furthermore, many teachers have not understood how to use the information which tests yielded. Hence the results of the tests appeared to have relatively little value. Recently many refinements of our educational tests have been introduced for the purpose of facilitating both the administration and the scoring of the test papers. It is now generally agreed that we have many educational tests which yield large returns upon the time and money invested in them. In the case of a battery of educational tests, such as the Illinois Examination or the Stanford Achievement Test, the investment of both the time and money is relatively large, but the returns are also large. One should never condemn a test as being too expensive without carefully considering the amount of information which it yields.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What meaning should be attached to the A.Q.?
2. Which furnishes the more satisfactory basis for school grades, A.Q. or achievement age? Why?
3. Why should a general intelligence test be included in a testing program?
4. Plan a testing program for the diagnosis of pupils in the operations of arithmetic.
5. Compare the Illinois Examinations with the Stanford Achievement Test. What are the particular merits of each?
6. What is a derived score?
7. Plan a testing program for the survey of a school system, including both elementary school and high school.

SELECTED BIBLIOGRAPHY

- Chapman, J. Crosby. "Convenience and Uniformity in Reporting Norms for School Tests"; in *Journal of Educational Research*, vol. 5, pp. 406-420. (May, 1922.)
- Franzen, Raymond. "The Accomplishment Quotient"; in *Teachers College Record*, vol. 21, pp. 432-40. (November, 1920.)
- Hull, Clark L. "The Conversion of Test Scores into Series which shall have any Assigned Mean and Degree of Dispersion"; in *Journal of Applied Psychology*, vol. 6, pp. 298-300. (September, 1922.)
- Knight, Frederick B., and Franzen, Raymond H. "Pitfalls in Rating Schemes"; in *Journal of Educational Psychology*, vol. 13, pp. 204-13. (April, 1922.)
- MacPhail, Andrew H. "The Correlation between the I.Q. and the A.Q."; in *School and Society*, vol. 16, pp. 586-88. (November 18, 1922.)
- McCall, William A. "Proposed Uniform Method of Scale Construction"; in *Teachers College Record*, vol. 22, pp. 31-51. (January, 1921.)
- Madsen, I. N. "The Correlation between Intelligence and Accomplishment Quotients"; in *School and Society*, vol. 16, pp. 696-97. (December 16, 1922.)
- Monroe, Walter S. *The Illinois Examination*. University of Illinois Bulletin, vol. 19, no. 9, Bureau of Educational Research Bulletin, no. 6. (Urbana: University of Illinois, 1921. 70 pp.)
- Murdoch, Katherine. "The Accomplishment Quotient — Finding and Using it"; in *Teachers College Record*, vol. 23, pp. 229-39. (May, 1922.)
- Otis, Arthur S. "The Method for Finding the Correspondence between Scores in Two Tests"; in *Journal of Educational Psychology*, vol. 13, pp. 529-45. (December, 1922.)
- Pintner, Rudolf, and Marshall, Helen. "A Combined Mental-Educational Survey"; in *Journal of Educational Psychology*, vol. 12, pp. 32-43. (January, 1921.)
- Pintner, Rudolf, and Fitzgerald, Florence. "An Educational Survey Test" in *Journal of Educational Psychology*, vol. 11, pp. 207-23. (April, 1920.)
- Pintner, Rudolf, and Marshall, Helen. "Results of the Combined Mental-Educational Survey Tests"; in *Journal of Educational Psychology*, vol. 12, pp. 82-91. (February, 1921.)
- Pressey, S. L. "Scale of Attainment No. 2 — An Examination for Measurement in History, Arithmetic, and English in the Eighth Grade"; in *Journal of Educational Research*, vol. 3, pp. 359-69. (May, 1921.)
- Pressey, Luella W. "Scale of Attainment No. 1 — An Examination of Achievement in the Second Grade"; in *Journal of Educational Research*, vol. 2, pp. 572-81. (September, 1920.)
- Pressey, Luella W. *Reading Scales for the Second, Third, and Fourth Grades*. University of Indiana Extension Division Bulletin, vol. 6, no. 12, pp. 46-52. (Bloomington: University of Indiana, 1921.)

- Pressey, Luella Cole. *The Relation of Intelligence to Achievement in the Second Grade*. Indiana University Extension Division Bulletin, vol. 6, no. 1, pp. 68-77. (Bloomington: University of Indiana, 1920.)
- Pressey, Luella C. "A First Report on Two Diagnostic Tests in Silent Reading for Grades II to IV"; in *Elementary School Journal*, vol. 22, pp. 204-11. (November, 1921.)
- Stebbins, Rena, and Pechstein, L. A. "Quotients I, E, and A"; in *Journal of Educational Psychology*, vol. 13, pp. 385-98. (October, 1922.)
- Thorndike, Edward L. "On Finding Equivalent Scores in Tests of Intelligence"; in *Journal of Applied Psychology*, vol. 6, pp. 29-33. (March, 1922.)
- Toops, Herbert A. "Determining Chronological Age in Decimal Parts of a Year"; in *Journal of Educational Research*, vol. 6, pp. 438-40. (December, 1922.)
- Toops, Herbert A., and Symonds, P. M. "What shall we Expect of the A.Q.?" in *Journal of Educational Psychology*, vol. 13, pp. 513-28; vol. 14, pp. 27-37. (December, 1922 and January, 1923.)
- Torgerson, T. L. "The Efficiency Quotient as a Measure of Achievement"; in *Journal of Educational Research*, vol. 6, pp. 25-32. (June, 1922.)

CHAPTER XI

THE CONSTRUCTION OF STANDARDIZED TESTS

It is the purpose of this chapter to give only a general description of the steps involved in the construction of a standardized educational test. For details, particularly statistical procedures, it will be necessary to consult another source.¹ However, it is believed that a general account of test construction will be helpful in understanding standardized tests and also in improving written examinations.

Value of a general knowledge of test construction. The principles of test construction afford a number of suggestions for improving the examination questions prepared by teachers and other school officials. From the same principles suggestions may be derived for the improvement of the administration of ordinary examinations and the grading of the examination papers. Another value grows out of the need for selecting tests for use. We now have a large number of standardized educational tests for most of the school subjects. There are a considerable number which have become widely used and new ones are being announced frequently. These tests are not all alike. They differ with respect to function as well as in a number of other ways.

¹ A complete account of test construction, including fundamental principles and statistical methods, is given in Monroe, Walter S., *An Introduction to the Theory of Educational Measurements* (Houghton Mifflin Company, 1923). A less comprehensive account may be found in McCall, W. A., *How to Measure in Education* (Macmillan Company, 1922). It will also be helpful to consult the accounts of the construction of particular tests, but when doing so it is necessary to bear in mind that such accounts describe the making of a particular test and are not comprehensive statements of procedure. Bibliographies will be found in the above sources, particularly the first, chapters IV, V, VI, and VII.

Some are power tests; others are designed to measure rate; some yield general measures; while others are to be used for the diagnosis of pupils with respect to their achievements. A few tests are designed to yield measures which are prognostic of a pupil's future success in a given subject. From this bewildering array of tests it is necessary for schoolmen to select those which are most nearly in accord with their purposes. Although the proof of a test is in its use, an understanding of how tests are made will be of great assistance in making a wise selection.

1. Principles of test construction

Determination of the minimum essentials of a subject. The first step in the making of a standardized educational test for measuring achievements of school children is the decision as to just what is to be measured. This is essentially the problem which is dealt with in curriculum construction, but it is a prerequisite step for test-making. For example, if we wish to construct an achievement test for measuring information in geography, it is necessary to determine the items of information which are to be included in the test. If we assume the purpose of constructing a test which will measure a pupil's acquaintance with the items of geographical information, which may be considered of fundamental importance, then we face the problem of determining what these items of information are. It is possible that when such a fundamental list of geographical information has been compiled, all of the items cannot be included in the test. It will be necessary, then, to make a selection from this number, but the content of the final test should be representative of this list.

In some fields the making of a test may not appear to have been preceded by any determination of minimum essentials of the subject. This is because the test-maker has assumed

that certain items of subject-matter are fundamental and has based his test upon these. For example, test-makers in the field of the fundamental operations of arithmetic have assumed that these operations with integers were legitimate educational objectives and have made them the basis of their tests. In most fields, however, the minimum essentials are not so obvious. In spelling, for example, it was not possible to construct satisfactory tests until Ayres determined the most commonly used words of the English language. With this list at hand, it is easy for test-makers to select words appropriate for the type of test they decide to construct.

Various methods of determining the minimum essentials of a given subject have been employed. Theoretically the procedures may be as varied as the procedures in curriculum construction. One of the most common methods employed by test-makers is to examine a number of textbooks on the given subject and to accept as fundamental those items which are common to these textbooks. In this way they secure the consensus of opinion of the authors who may be regarded as experts in this field. Furthermore, they secure a list of items which may be considered to be generally taught. This method has certain obvious limitations, but it has been used because of the accessibility of the material and the routine character of the procedure.

The determination of the minimum essentials of a school subject is the basic and fundamental step in the construction of a standardized educational test. Some test-makers do not appear to have taken this step seriously, and as a result their tests are open to criticism on the ground that they include exercises which are relatively unimportant when judged by our educational objectives.

Type of test to be constructed. From the standpoint of structure our educational measuring instruments may be classified under three heads,

1. *Rate tests.* Rate tests usually consist of exercises approximately equal in difficulty, and the time allowed is such that no pupil can finish the test. The number of exercises done is taken as a measure of the pupil's rate of work. The quality of his performance may be expressed by the per cent of exercises done correctly. Sometimes rate and quality are combined by taking as a pupil's score the number of exercises done correctly. The following are among the most widely used rate tests: Curtis Standard Research Test in Arithmetic, Series B; Monroe Standardized Silent Reading Test, Revised; The Cleveland Survey Arithmetic Tests; Burgess Picture Supplement Scale for Measuring Reading Ability; and the Curtis Silent Reading Test No. 2.

2. *Power tests.* A power test consists of a series of exercises within a given field of subject-matter arranged in ascending order of difficulty, and is intended to measure a pupil's power to do more and more difficult exercises of the type included in the test. This type of measuring instrument is frequently called a "scale." No measure of a pupil's rate of work is usually secured because the time allowed is sufficient for all to complete as many of the exercises as they are able to do. Theoretically, a pupil's score on such a test is the degree of difficulty of the most difficult exercises he is able to do with a specified degree of accuracy. Such a score is laborious to compute from a pupil's performance, and for this reason the number of exercises done correctly is generally taken as his score. Examples of this type of measuring instrument are: Woody Arithmetic Exercises, Thorndike-McCall Reading Scale, Hotz Algebra Scales, Henmon Latin Tests, and Van Wagenen Reading Scales.

3. *Quality Scales.* In such subject-matter fields as handwriting, English composition, handsewing, etc., the measuring instruments are of a different type. In these subjects the pupil produces a performance which cannot be con-

sidered as either wholly right or wholly wrong. The problem is to describe its degree of quality. For this purpose "quality scales" have been constructed.¹ These consist of a series of specimen performances arranged in order of ascending merit. In measuring achievement in such a subject the pupils are asked to produce a sample performance under specified conditions. This is then described by comparison with the quality scale.² The following quality scales have been widely used: Thorndike Handwriting Scale; Ayres Handwriting Scale, Gettysburg Edition; Willing Scale for Measuring Written Composition; and Murdoch's Scale for Measuring Hand Sewing.

In subjects where the test consists of exercises which the student is asked to do, a choice must be made between a power test and a rate test. These two types of tests measure different kinds of mental growth. A rate test measures a pupil's skill in doing a certain type of exercise. Usually the exercises are relatively easy, and there is little or no question about a pupil being able to do them. Such a test implies that our educational objectives are to train children to be more and more skillful or fluent in doing the particular type of exercise included in the test. On the other hand, a power test, or "scale," measures a pupil's ability to do more and more difficult exercises within a given field of subject-matter. Thus it is implied that our purpose is to train children to do more and more difficult exercises of the kind represented in the test. Up to a certain point this is in agreement with our general educational objectives, but no one would argue that we should aim to train pupils to do exercises simply because they are difficult. For example, in spelling we teach pupils to spell words, not because they are

¹ The construction of a "quality scale" is not described in this chapter.

² The rate of work may also be measured, but this is done independently of the particular scale used.

difficult, but because the pupil will use the words in his written language. Similarly, in arithmetic we train pupils to do certain types of examples because similar examples will be encountered in the solving of problems. No type of example is given a place in the list of objectives in arithmetic simply because it possesses a certain degree of difficulty. In fact, many exercises would be found difficult merely because they have been excluded from our list of minimum essentials and have not been taught to the pupils. Both power tests and rate tests have merit. For certain purposes a rate test is to be preferred; in other situations a power test will be more useful. The maker of a test should consider carefully which type of mental growth he desires to measure, and construct the test which will best fulfill his purpose.

Types of exercises. Achievement is measured through performance. In a test a pupil is asked to work an example, to solve a problem, to answer a question, or to do something else which will yield a performance satisfactory for testing purposes. In deciding upon the type of exercise to be used, there are several criteria which should be kept in mind. In the first place, the exercises should be of such a nature that the pupil will be called upon to do relatively little writing. This is especially important in the case of rate tests. In the answering of ordinary questions a pupil's response is influenced by his rate of writing and by his ability to express his ideas. Thus, the score which is given to a paper will depend in part upon the ability to write and the ability to express ideas.

In order to have a test highly reliable it is necessary that the exercises call for a response which may be judged as either right or wrong and which will be judged in the same way by all competent persons. When the description of a student's answer is a matter of opinion, scorers will assign different marks to it. This is one of the reasons why ordinary examinations yield inaccurate measures of achieve-

ment. Hence, so far as is compatible with other requirements, exercises should be used for which there is one and only one correct answer.

Finally, it is desirable that the responses called for by an achievement test approach as nearly as possible typical classroom performances. In arithmetic pupils are asked to do examples as regular school exercises. In an arithmetic test consisting of examples the pupils are asked to do exactly the sort of thing they have been trained to do. Furthermore, arithmetic examples permit of only one correct answer and require relatively little writing on the part of the pupil. In other subject-matter fields many typical classroom exercises are unsatisfactory for testing purposes. Some of them require a great deal of writing on the part of the student and others do not permit of objective scoring. Hence, it has been necessary for test-makers to construct exercises for testing purposes which in many cases the pupils have not met in their regular class work.

Test-makers have exhibited much ingenuity in devising exercises which are suitable for testing purposes. In Monroe's Standardized Silent Reading Test, Revised, the student is asked to underline or indicate in some other way the one word, in a list of five, which best answers the question asked. Courtis has devised a geography test in which a pupil merely has to write certain numbers. In Barr's Diagnostic Test in American History a pupil gives most of his responses by checking certain items. The same procedure is followed in the Haggerty Reading Examination, Sigma 3. In the Charters Diagnostic Language and Grammar Test the pupil indicates the grammatical rule which applies to a given sentence by means of writing a number. In the case of all these tests one and only one response is correct. Thus there can be no difference of opinion concerning the score which is given a pupil on such a test.

Selection of exercises for the final form of a test. In the construction of a standardized educational test an important step is the selection of the exercises for the final form of the test. Even when great care has been taken in the formulation of the exercises, some of them frequently are found unsuitable for testing purposes. Thus, after a tentative test has been constructed it should be given to a considerable number of children in each of the grades for which it is intended. On the basis of this trial the exercises for the final form of the test should be selected. It may even be necessary to add exercises before a satisfactory test is secured. Any exercise which is not interpreted in the same way by all pupils should be rejected. Exercises which are difficult to score should be omitted when possible. A further selection should be made with respect to the particular type of test being constructed. In the case of a rate test the exercises should be as nearly uniform in difficulty as possible. Unusually easy ones or exceptionally difficult ones should not be included. In a power test there should be some exercises so easy that they will be done correctly by practically all children, and some so difficult that few, if any, children can do them. All levels of difficulty between these extremes should be represented. In constructing a power test it is customary to reject exercises which are not done correctly by a gradually increasing per cent of children as one passes from a lower to a higher school grade. For example, any exercise which is done correctly by a smaller per cent of children in the fifth grade than in the fourth would be rejected as being unsatisfactory for testing purposes. In view of the very obvious influence of instruction upon the difficulty of an exercise, one might raise the question of the soundness of this criterion until such time as our courses of study may be considered ideal. It is also highly desirable to select only those exercises for the final form of a power

test which will form a gradually increasing scale of difficulty from the easiest to the most difficult. In order to satisfy this criterion, it is sometimes necessary to formulate new exercises and determine their difficulty.

Directions for administering standardized tests. The directions for administering a standardized educational test and for scoring the papers form an essential part of the measuring instrument. The norms which are announced for a test apply only when the test is used under standard (uniform) conditions, and obviously these conditions cannot be realized unless the test is accompanied by detailed directions. Unless the test consists of exercises with which pupils have become acquainted in a classroom, as in the case of arithmetic, the nature of the exercises which the pupils are to do should be carefully explained to them. It is well to give them a little practice in doing such exercises before they begin the test proper. Furthermore, they should be given precise directions concerning what to do, what not to do, how rapidly to work, and how long they will have for the test. Unless the scoring of the test is objective, detailed instructions must be given for this also. Usually a standardized educational test is accompanied by a class record sheet in order to facilitate the assembling of the scores of a class and to assist in the interpretation of these scores.

Standardization of a test. The standardization of a test refers to the deriving of norms for interpreting the scores yielded by it and not to any characteristic of the form of the test. Some users of educational tests appear to have considered "standardized" to mean that the score was expressed in terms of a "standard unit." In the case of physical measurements we have standard units and a standard measuring instrument is one in which the standard unit of the dimension measured is incorporated. In the field of mental measurements the term "standard" has an entirely different meaning.

The method of standardizing a test is to have it given to a large number of representative pupils in each of the grades or age groups for which it is intended to be used. In order to insure that these pupils are representative, the test is given in several cities and usually in several different sections of the country. From scores secured in this way two types of norms have been derived. The usual procedure is to assemble the scores by grades and to calculate the median score for each grade. These median scores form "grade norms." "Age norms" are secured by assembling the scores on the basis of the chronological ages of the pupils; that is, the scores of all ten-year-old pupils would be brought together in one distribution, the scores of eleven-year-old pupils in another, and so on. The medians of these distributions constitute the age norms. It is obvious that the securing of age norms is somewhat more difficult than the procedure for grade norms. In the first place, it is more difficult to assemble the scores on the basis of age than of grade, and, in the second place, it is somewhat difficult to secure scores from all pupils of a given age. For example, if we wish to secure an age norm for twelve years, some twelve-year-old children will be found in the high school and in each of the four or five grades immediately below the high school. A few are likely to be found in the primary grades.

Grade norms are somewhat unsatisfactory because a given grade in one school does not mean necessarily the same thing as the same grade in another school. The policy of promotion which prevails may result in the children being chronologically several months older in the fifth grade in one school system than those in the fifth grade of another. For this reason, it has been urged that age norms are much more desirable, since a pupil's chronological age is definite and is not influenced in any way by the plan of promotion or the general organization of the school.

It is generally thought that the achievements of pupils in rural schools are sufficiently different from those of pupils in urban schools to warrant the determination of separate norms for each of these types of schools. Some have urged even the desirability of making a distinction between large and small cities. Since a pupil gradually increases in achievement throughout the school year, it is desirable to have the norms stated for a specific month. From the norms stated for a given month of the school year it is possible to estimate the norms for any other month.

Derived scores. The construction of many recent tests has included the determination of the basis for translating the point scores into derived scores. The various types of derived scores have been described on page 380. This is not an essential step, but it adds greatly to the usefulness of the test for certain purposes.

Critical evaluation of a test. The final step in test-making is a critical evaluation of the test which has been made. Perhaps the most important single item of information to be secured under this head is the determination of the reliability of the test. This is accomplished by administering the test twice to the same group of pupils and calculating the coefficient of correlation between the two sets of scores. It is desirable that different forms of the test be used for this purpose and that the two administrations of the test be separated by a relatively short time interval. The coefficient of reliability is a measure of the extent to which the scores obtained from a second application of the test will agree with those obtained from the first application. It is an index of the variable errors of measurement. A more easily interpreted index of these errors is the probable error of measurement. (See page 408.)

A critical study of a standardized educational test should include also comparisons of the measures obtained with

those yielded by other tests having a similar function and with other criteria whenever possible. In case duplicate forms have been constructed there should be a determination of the degree of equivalence of these forms. It is also important to know the practice effect which may be expected when the test is repeated after a short interval of time, even though a duplicate form is used.

In addition to such critical studies of a test which are largely statistical, a careful analysis of the content and structure of the test is frequently helpful in determining its value. The content of a test should be as nearly as possible in agreement with recognized educational objectives. The results obtained by using a test are also indicative of its value. As indicated in the beginning, the proof of the test is in its use. That test which proves most helpful in realizing a particular purpose is the best test for that purpose.

2. An illustration of the critical evaluation of a test

In order to illustrate the procedure of making a critical study of a test, a portion of the account of the derivation of the Illinois Examination ¹ (see page 382) is reproduced here. The first three divisions of the account, which are omitted, include: (1) a statement of facts of title; (2) a description of the types of exercises and facts concerning equivalence of the duplicate forms; (3) an account of the description of a pupil's performance, including the calculation of derived scores. We begin with IV. Function.

IV. FUNCTION

The function of the three scales which make up the Illinois Examination is implied in their structure. The Illinois General

¹ This account is taken with a few minor changes from Monroe, Walter S., *The Illinois Examination*, Bureau of Educational Research Bulletin, no. 6, University of Illinois Bulletin, vol. XIX, no. 9, October 31, 1921.

Intelligence Scale provides a measure of general intelligence of children in Grades III to VIII, inclusive. Monroe's Standardized Silent Reading Tests, Revised, are intended to yield measures of the ability to read silently simple descriptive and narrative material when the reading is done for the purpose of answering questions. Monroe's General Survey Scale in Arithmetic is designed to yield general measures of a pupil's ability to perform the operations of arithmetic. It should be noted that the function of these tests is general rather than diagnostic. It is possible to use the sub-tests of the General Survey Scale in Arithmetic as diagnostic tests although they were not designed for this purpose.

V. VALIDITY

The ideal procedure to be followed in studying the truthfulness of the measures yielded by the scales included in the Illinois Examination would be to compare them with true measures secured by other means. However, in no case are such true measures available. It is, therefore, necessary to study the validity of these scales by methods which are obviously imperfect.

1. **Objectivity.** The scales of the Illinois Examination are highly objective with respect to the scoring of test papers. Except when a pupil fails to follow directions no questions, concerning which answers are correct, arise. The administration of the tests is also highly objective. The directions for examiners have been found to be adequate and in all cases the examiner is told very explicitly what he is to say to the pupils. Much of the explanation is also printed on the test booklet so that the pupil has an opportunity to read as well as to hear the explanation.

2. **Reliability.** In order to study the reliability of the three scales which make up the Illinois Examination the different forms were given to the same pupils. The instruction to those coöperating in this study was to give all the forms within the same half-day. The scores on the different forms were compared by means of the Pearsonian coefficient of correlation and by means of other statistical devices which will be explained in the following pages.

The coefficient of correlation merely indicates the relationship between two sets of scores. It is simply an index of the extent to which the pupils make the same score on the second trial of a test that they make upon the first trial when the practice effect is disregarded. In Figure 18, we represent graphically the scores made by

the fifth grade pupils on Forms 1 and 2 of the Illinois General Intelligence Scale. The coefficient of correlation (r_{12}) for this group of scores is $.92 \pm .006$. It is obvious that in some cases the pupils make the same score or approximately the same score on second trial. In other cases there are marked differences between the scores on the two trials.

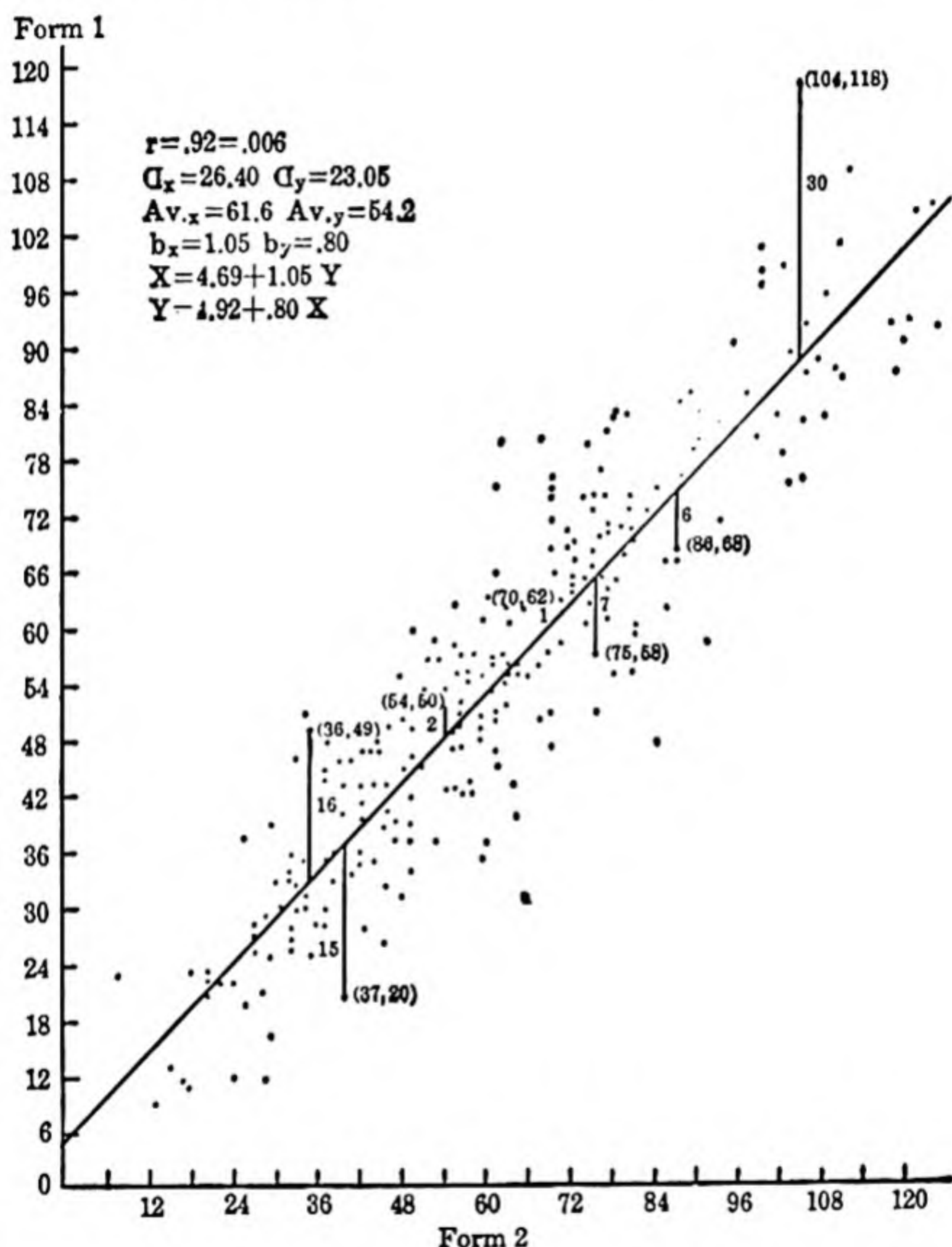


FIG. 18. CORRELATION OF FORM 1 SCORES WITH FORM 2 SCORES OF THE ILLINOIS GENERAL INTELLIGENCE SCALE, FIFTH GRADE

In Figure 18 the regression line, $y = 4.92 + .80 x$, has been drawn. Perfect correlation ($r_{12} = 1.00$) would be secured if the Form 1 scores were changed so that all points would fall upon this

regression line. This would require a vertical shifting of the points. Those above would be moved downward, while those below would be moved upward. For a few of the points vertical lines have been drawn in to indicate the amount of shifting necessary. Perfect correlation would be secured if this were done with reference to any line, but this regression line is the one for which the standard deviation of the shifting is the least. The other regression equation, $x = 4.69 + 1.05 y$, possesses similar properties for a horizontal shifting.

The amounts of change necessary to secure perfect correlation may be thought of as departures from perfect correlation. The magnitude of these changes is described by the equation for the probable error of estimate,

$$\text{P.E.}_{\text{est}} = .6745 \sigma_y \sqrt{1 - r_{12}^2}$$

Substituting in this equation ¹ for σ_y and r_{12} we have the probable error of estimate equal to 6.06. The probable error of estimate is more easily interpreted as the index of the degree of correlation that exists than the coefficient of correlation.

Since neither set of scores are accurate measures of intelligence the differences between the pairs of scores do not truthfully represent the degree of inaccuracy of either set of scores. The error of any score is the difference between it and a pupil's true score. We may define a true score as the average of an infinite number of scores after they have been corrected for practice effect, fatigue, and other factors which would tend to increase or decrease the averages of the successive sets of scores. Such true scores are obviously not obtainable. It is, however, possible to determine the coefficient of correlation between either set of obtained scores and the corresponding true scores. This is done by the formula,²

$$r_{1t} = \sqrt{r_{12}}$$

In this formula r_{12} is the coefficient of correlation between the two sets of obtained scores and r_{1t} is the coefficient of correlation between one set of obtained scores and the corresponding true scores. To distinguish the coefficient of correlation of a set of obtained

¹ The table which gives σ_y and r_{12} is not reproduced.

² See Kelley, T. L., "A Simplified Method of Using Scaled Data for Purposes of Testing"; in *School and Society*, vol. 4, p. 74 (July 8, 1916), and "The Reliability of Test-Scores"; in *Journal of Educational Research*, vol. 3, pp. 370-79. (May, 1921.)

scores with the corresponding set of true scores from the coefficient of correlation between two sets of obtained scores we call the latter the *coefficient* of reliability and the former the *index* of reliability.

Table XXVIII gives coefficients of reliability (r_{12}) and the indices of reliability (r_{1t}) for each of the three scales which make up the Illinois Examination. It will be noted that a high degree of reliability is indicated in most cases. In some instances it is unusually high in comparison with the degree of reliability reported for other tests. In the case of silent reading certain instructions were not followed by some of the examiners, and it is thought that their failure to do so caused the two sets of scores to correlate less highly than they should.

Both the coefficient of reliability and the index of reliability are difficult to interpret. They express a general relationship but not in terms of the actual amount of error which must be allowed for in the case of the scores of individual pupils. It is possible to calculate another and more easily interpreted expression of the reliability or accuracy of the measures yielded by tests. The probable error of estimate is given by the formula,

$$\text{P.E.}_{\text{est}} = .6745 \sigma \sqrt{1 - r_{1t}^2}$$

In the formula σ may be taken as either σ_1 or σ_2 . Theoretically, these are expected to be equivalent. Practically, slight differences may exist. It is, therefore, advisable to use as the value of σ the average of σ_1 and σ_2 . The probable error of estimate (P.E._{est}) in this case is the probable error of the measurements yielded by the test. Hence, we call it by this name and use the symbol, P.E._m . Since $r_{1t} = \sqrt{r_{12}}$, the above formula may be written in the form,

$$\text{P.E.}_m = .6745 \sigma \sqrt{1 - r_{12}}$$

The probable error of measurement calculated by the above formula is to be interpreted as an index of the amount of departure of the obtained scores from the true scores. In other words, it is the error which the obtained score involves. This error is described as a probable error. Such a description, of course, tells us nothing about the magnitude of the error in the case of a particular pupil but it does describe in a general way the magnitude of the errors involved in a group of scores.¹

¹ The probable error of measurement refers to only the variable errors. It is not influenced by the constant errors.

TABLE XXVIII. RELIABILITY COEFFICIENTS

GRADE	GENERAL INTELLIGENCE			ARITHMETIC			SILENT READING				
	No. of pupils	r_{12}	r_{14}	No. of pupils	r_{12}	r_{14}	No. of pupils	Rate		Comprehension	
								r_{12}	r_{14}	r_{12}	r_{14}
III.....	76	.86	.93	229	.80	.89	116	.69	.83	.63	.79
IV.....	120	.93	.96	271	.86	.93	112	.79	.89	.63	.79
V.....	243	.92	.96	256	.88	.94	120	.79	.89	.69	.83
III to V....				826	.95	.99	348	.78	.88	.80	.89
VI.....	198	.82	.91	271	.76	.87	139	.87	.93	.52	.72
VII.....	157	.80	.89	257	.71	.84	100	.72	.85	.68	.82
VIII.....	164	.67	.82	171	.79	.89	119	.91	.95	.85	.92
VI to VIII				699	.76	.87	358	.79	.89	.72	.85
III to VIII	958	.92	.96								

The probable error of measurement tends to increase as the scores become larger and the significance of an error depends upon the magnitude of the score with which it is associated. For this reason added meaning can be given to our description of the errors by calculating the ratio of the probable error of measurement to the average score. This gives the probable error of measurement in the form of a per cent of the score. In Table XXIX, the probable errors of measurement and the ratios of these to the average scores are given. The numbers of pupils involved are the same as those given in Table XXVIII. A probable error of measurement of 3.5 for the Illinois General Intelligence Scale in the third grade means that the point scores of fifty per cent of the pupils will involve errors less than this amount. The remaining fifty per cent of the scores will involve errors greater than 3.5. A somewhat more general statement is that, on the average, the scores obtained from third grade pupils will involve probable errors of ten per cent of their magnitude. In the case of the Illinois General Intelligence Scale the average probable error of measurement amounts to about six months. This is approximately the same as that cal-

TABLE XXIX. PROBABLE ERRORS OF MEASUREMENT AND RATIO OF PROBABLE ERRORS OF MEASUREMENT TO AVERAGE SCORES

GRADE	INTELLIGENCE		ARITHMETIC		SILENT READING			
					Comprehension		Rate	
	P.E. _m	P.E. _m	P.E. _m	P.E. _m	P.E. _m	P.E. _m	P.E. _m	P.E. _m
		Av.		Av.		Av.		Av.
III.....	3.5	0.10	2.6	0.17	1.2	0.16	13.7	0.12
IV.....	5.5	0.09	4.6	0.12	1.4	0.16	10.3	0.08
V.	4.7	0.08	4.4	0.09	1.2	0.10	12.0	0.08
III to V...			3.2	0.10	1.0	0.11	13.1	0.10
VI.....	5.5	0.07	6.3	0.10	1.3	0.10	9.1	0.05
VII.....	6.4	0.07	5.3	0.09	1.2	0.09	13.6	0.07
VIII.....	7.7	0.08	5.4	0.08	0.8	0.05	7.5	0.04
VI to VIII.			6.2	0.10	1.1	0.08	12.0	0.07
III to VIII.	5.3	0.07						

culated for the Stanford Revision of the Binet Scale for the Measurement of Intelligence.¹

3. Discrimination. The shape of the distribution of the scores which a test yields throws some light upon its validity. In order that a test be valid the scores must show differences of the traits measured when these differences exist. When a representative group of pupils is measured with reference to a mental or a physical trait we may expect to find a distribution closely approximating the normal shape. When the number of cases is large this approximation should be close if there is proper discrimination. Any marked departure from the normal shape indicates that for some pupils at least there is a lack of discrimination. On the other hand, when we have a normal distribution we cannot know defi-

¹ Otis, A. S. and Knollin, H. E. "Reliability of Binet Scale and Pedagogical Scales"; in *Journal of Educational Research*, vol. 4, pp. 121-43. (September, 1921.)

nately that our measures are accurate. The shape of the distribution, therefore, has only a negative significance. We can only say that when there is a striking departure from normality there is a lack of discrimination and, hence, inaccuracy of measurement for some pupils.¹

4. **Comparison with criterion measures.** The Illinois General Intelligence Scale was given to 203 pupils whose mental ages had also been determined by the Stanford Revision of the Binet Scale for Measuring Intelligence. The correlation between the mental ages as determined by these two scales is $.74 \pm .02$. The probable error of estimate is 1.2 years. This means that in 50 per cent of the cases, the mental age, as determined by the Illinois General Intelligence Scale, differed from that as determined by the Binet Scale by 1.2 years or less. This lack of agreement between the measures secured by these two scales is not due solely to errors in the measures yielded by the Illinois General Intelligence Scale. The Stanford Revision of the Binet Scale for Measuring Intelligence also yields measures which involve errors of measurement of approximately the same magnitude as those of the Illinois General Intelligence Scale.

In November, 1920, both the Illinois General Intelligence Scale, Form 1, and the National Intelligence Scale, Form 1, were given to 3615 pupils in eight elementary schools in Chicago. The correlation between the scores obtained from these two tests is indicated in Table XXX. The probable errors of estimate indicate that the agreement is not close. The probable error of estimate, when all grades are taken together, is 11.5. This means that the departure from perfect correlation with the scores yielded by the National Intelligence Scale is greater than 11.5 points in 50 per cent of the cases, and less than 11.5 in 50 per cent of the cases. Since ten points are equivalent to one year of mental age, the relationship between the scores yielded by the Illinois General Intelligence Scale and by the National Intelligence Scale is approximately the same as the relationship shown to exist between the scores yielded by the Illinois General Intelligence Scale and the Stanford Revision of the Binet Scale for Measuring Intelligence.

The Illinois General Intelligence Scale, Form 1, was given to a number of sixth-grade pupils whose I. Q.'s, as determined for the

¹ The consideration of the "discrimination" of the Illinois Examination is omitted because of the lack of space.

TABLE XXX. CORRELATION BETWEEN SCORES YIELDED BY ILLINOIS GENERAL INTELLIGENCE SCALE AND BY NATIONAL INTELLIGENCE SCALE

GRADE	NUMBER OF CASES	r	P.E.est	P.E.est
				Av.
III A....	357	0.53	9.1	0.22
IV B....	416	0.70	9.6	0.18
IV A....	335	0.74	8.0	0.14
V B....	460	0.55	8.7	0.14
V A....	285	0.47	12.0	0.19
VI B....	383	0.44	12.6	0.17
VI A....	259	0.67	10.8	0.13
VII B....	350	0.70	11.0	0.12
VII A....	210	0.68	10.3	0.11
VIII B....	271	0.72	10.2	0.10
VIII A....	289	0.69	10.9	0.10
All Grades	3615	0.81	11.5	0.16

Otis Group Intelligence Tests, were available.¹ The coefficient of correlation for 83 VIA pupils was $.82 \pm .02$. For 124 VIB pupils the value of r was $.83 \pm .02$. The probable error of estimate was 6.4 in the first case and 5.9 in the latter.

The Pintner Non-Language Group Intelligence Tests are represented to have a reliability coefficient of .72. This is by mental indices and not by point scores. The two sets of measures were obtained by use of the same test after an interval of two years. The number of children tested was 46. These group intelligence tests were also given during the same semester to 300 children whose mental ages had been determined by the Stanford Revision of the Binet Scale for Measuring Intelligence. The coefficient of correlation between the point scores yielded by the Pintner Group Test and the mental ages determined by the Binet Test was .80.

No data are available at this time for making comparison between the measures of achievement yielded by the achievement

¹ The writer is indebted to Superintendent L. W. Keeler, Michigan City, Indiana for these data.

scales included in the Illinois Examination and by other similar scales. Neither are data available for comparison of measures of achievement with teachers' estimates.

5. **Inferences concerning validity based upon the structure of the test and its administration.** In the case of the Illinois General Intelligence Scale the sub-tests have frequently been used by other makers of instruments for measuring general intelligence. At the time the Illinois General Intelligence Scale was constructed a number of other intelligence scales were analyzed with reference to sub-tests and the ones most frequently found were incorporated in this scale.

The Illinois General Intelligence Scale is explicitly a verbal test. Ability to read is a prerequisite. For this reason it may be urged that it does not permit non-verbal elements of intelligence to function. This objection, of course, applies to other verbal tests. No data are at hand to show the limitation which this feature places upon the scale.

The silent reading test included is a revision of the Monroe Standardized Silent Reading Test. In this revised form certain features of the original test which were found unsatisfactory have been eliminated. The scoring has been made objective. The exercises are more uniform and a more precise measurement of the rate of silent reading is secured. Experience has shown that the test is too short for the time limit allowed. This, in the case of the most fluent readers, prevents one from securing valid measures of reading ability. To one acquainted with the nature of reading ability and its measurement, this one test obviously measures in a general way only one type of silent reading ability. To measure completely all phases of silent reading ability would require a battery of tests.

In the case of the Monroe General Survey Scale in Arithmetic the sub-tests are judged to represent the most important types of examples learned by pupils in the sequence of grades for which they are intended. The choice represents the judgment of the author but is not inconsistent with other groups of tests used to measure ability of pupils in the operations of arithmetic. A single general score has been used to describe the pupil's ability. This is, obviously, a composite not only of the scores of the different sub-tests but also of the dimensions of rate and accuracy. However, since a single general measure is desired this does not constitute a serious criticism of the scale.

Summary for validity. By way of summary we may say that the scales which make up the Illinois Examination compare favorably in respect to validity with our best tests. It is, however, clear that the scales possess certain limitations which should be kept in mind when the scores are interpreted.

7. Norms: Practice effect when a test is repeated.¹ The grade norms obtained are for the first application of the Illinois Examination. When it is given a second time pupils will tend to make higher scores because of their acquaintance with the nature of the tests. The amount of increase varies. If pupils are "coached" upon the tests a large increase is to be expected. When a period of several months intervenes between the first and second trials and the pupils have received no training upon the exercises of the tests, the increase appears to be small and in some cases can be neglected without serious error. No evidence was obtained relative to the effect of using the same form instead of a different form.

In order to ascertain the effect of practice when the second application immediately follows the first, both Form 1 and Form 2 were given to a number of pupils. Due to a miscarriage of plans the practice effect for the silent reading tests was not determined. For the Illinois General Intelligence Scale the average practice effect is approximately 5.0 points, or six months of mental age, if the returns from the eighth-grade pupils are not used. In this grade unusual conditions appear to have prevailed and when the scores from it are included the practice effect is approximately 7.0 points. For Monroe's General Survey Scale in Arithmetic the average practice effect is approximately 3.2 points in Grades III to V and 4.5 points in Grades VI to VIII. It is, therefore, obvious that when the Illinois Examination is repeated after only a short interval the second trial scores must be corrected before they can be compared with those obtained from the first trial.

In one school the teachers of 134 pupils gave special drill and instruction to their pupils after Form 1 of the Illinois Examination had been given in November. The teachers did not know that Form 2 was to be given later, and did not have in mind, therefore, preparing the pupils for it. Their instruction was, to a slight extent, based upon Form 1 test papers. Believing that their pupils were rather weak in knowledge of vocabulary and in synonym-

¹ Section 6, Validity of Significance, is omitted and the treatment of norms is abbreviated.

antonym, special drill was given along these lines in language work. In arithmetic practice was given upon those combinations where the pupils seemed weak. In reading there was some special drill for increasing the rate of silent reading. The gains made by the pupils under these teachers are given in Table XXXI.

TABLE XXXI. GAINS DUE TO SPECIAL INSTRUCTION UPON THE ILLINOIS EXAMINATION

	MEDIAN POINT SCORES			MEDIAN QUOTIENTS		
	Nov. 1920	May 1921	Gain	Nov. 1920	May 1921	Gain
Intelligence.....	57.6	100.8	43.2	99.6	128.0	28.4
Arithmetic.....	48.2	120.0	71.8	114.0	137.4	23.4
Comprehension..	10.9	15.5	4.6	105.1	102.0	-3.1
Rate.....	171.4	237.9	66.5	122.0	139.4	17.4

It is commonly assumed that general intelligence is unaffected by school instruction. A median gain of 43.2 points or slightly more than four years in mental age within a period of six months indicates that mental age, as measured by such an instrument as the Illinois Intelligence Scale, is affected by classroom instruction. It is, therefore, necessary to exercise a great deal of caution in interpreting changes in the intelligence scores derived from two successive testings separated by a considerable time interval. Even in the case of a first trial the scores obtained will be misleading if the pupils have received any special preparation for the test.

One would naturally expect that the achievement scores in arithmetic and silent reading would be materially affected by instruction. Table XXXI shows very large gains in achievement. The increases in achievement quotients show that except for the comprehension of silent reading the gains are relatively larger in achievement than in mental age. It is not unlikely that some of the increases in the May scores over the November scores are due to the pupils being more familiar with the testing procedure. The effect of familiarity was not investigated but due caution should be

exercised in interpreting the gains in achievement as being the result of instruction directed to the needs of the pupils.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. Define in your own words: objective, coefficient of correlation, probable error of estimate, probable error of measurement, educational objective, norm, derived score, educational age, and achievement quotient.
2. Examine the accounts of the construction and use of the Courtis Standardized Research Tests in Arithmetic, Series B, for the purpose of arriving at an estimate of the reliability and validity of these tests. Add to the experimental evidence any inferences which can be made and an analysis of the tests. Prepare a list of other items of information needed for a complete determination of the validity of this series of tests. (See Bibliography at end of Chapter II for references.)
3. Repeat this exercise with two other tests in which you are interested.

BIBLIOGRAPHY

For references to accounts of the construction of standardized tests and also for critical studies of tests, consult the bibliographies for Chapters II to IX. For a selected list of typical references, see bibliographies in Chapters IV, V, VI, and VII in Monroe, Walter S., *An Introduction to the Theory of Educational Measurements*. (Houghton Mifflin Company, 1923.)

CHAPTER XII

THE MEANING OF SCORES

Two methods of describing achievement. Until the advent of standardized educational tests, measures of achievements of school children resulting from written examinations and teachers' estimates were expressed in terms of such words and symbols as "fair," "good," "excellent," "A," "B," "C," "78%," and "95%." These measures are commonly called school marks or "grades." The measures of achievement yielded by standardized educational tests are generally called "scores" or "point scores." A pupil's examination paper may also be described in terms of a point score. He may receive a credit of 8 on the first question, 13 on the second, 18 on the third, 5 on the fourth, etc. The total number of points constitutes a description of his performance on the examination. When an examination paper is marked in terms of per cent, the "grade" is essentially a point score. The fact that the maximum number of points which may be received is 100 does not change the nature of the score.

A score is an absolute measure of achievement. Until it is compared with a norm, it has no meaning. A school mark or "grade" is an interpreted measure of achievement. It expresses a comparison of a measure of achievement with a norm. Usually this norm is subjective. It exists only in the mind of the person who gives the mark, and therefore different teachers would use different norms. The same teacher would likely use different norms at different times. A grade of "superior" in reading given to a fourth-grade pupil means that in the teacher's judgment he possesses a degree

of ability distinctly greater than the teacher's norm for fourth-grade pupils. If the pupil had been judged by another teacher who had in mind a different norm, the "grade" would have been different. Failure to recognize this distinction between scores and school marks constitutes one of the limitations of written examinations. (See page 7.)

Need for translating scores into school marks. Both scores and school marks possess certain limitations. Unless expressed with reference to a definite norm a school mark is indefinite. A given school mark is used with a different meaning at different times. Such words as "fair," "good," "excellent," describe certain degrees of ability in the case of a third-grade pupil and distinctly different degrees of ability in the case of fourth- and fifth-grade pupils. To define a grade of "excellent," or "A," as 95 to 100 per cent is merely to substitute one descriptive term for another. This does not tell us what the school mark means. On the other hand, to say that a pupil's point score on an examination is 41, or that his silent reading rate is 120, or that the number of examples done correctly on the Courtis Standard Research Test in Arithmetic, Series B, is 14, lacks meaning until a comparison is made with an appropriate norm. We have indicated one phase of the interpretation of test scores in the preceding chapters. The calculation of certain derived scores frequently facilitates this interpretation. However, for certain purposes it is desirable to translate point scores on examinations and test scores into school marks or "grades."

Before considering the technique of translating scores into school marks, we shall discuss the basis of satisfactory norms. As we have already indicated, the usual method of standardizing an educational test is to have it given to a large number of representative groups of pupils. The average or median score of a group is taken as the norm.

This procedure implies that on the average present conditions are satisfactory. If many of the critics of our schools are right in their contentions, present conditions are far from what they should be. Therefore, it is pertinent to inquire concerning the basis of satisfactory norms or standards.

A reasonable standard. A satisfactory standard must be reasonable and must be "efficient." To be reasonable a standard must be such that it can be attained by pupils, under school conditions, and with an appropriate time expenditure. Pupils are limited in their learning by inherited characteristics, and all cannot attain to the same levels of skill. However, it must not be forgotten that the medians, or averages, of present attainments of pupils are far below the levels of skill which have been attained by many pupils and adults. For example, the eighth-grade standard for the Courtis addition test is only eleven examples attempted. Some eighth-grade pupils are able to do twenty or more examples, and adults who have been specially trained have done fifty to sixty examples. In fact Courtis, who has studied arithmetical abilities for several years, states that it appears that there is no limit to the rate at which columns of figures may be added, provided the amount of time for practice is unlimited. In the school the time for practice is limited, but even then the standard of eleven examples is markedly below the attainment of many pupils under present school conditions, and it still remains to be seen what degree of ability to perform the fundamental operations of arithmetic with integers might be engendered with the present time allotment, provided the methods and devices of instruction were properly suited to the pupils. It seems probable that the standards which we have now are below the level of the possible attainment of a large per cent of pupils under school conditions. Just how large this per cent

will be when the methods and devices of instruction are appropriately adjusted to the pupils can be determined only by actual trial. It may be that, when we learn to adjust our methods and devices of instruction better to the abilities of pupils, a higher standard of ability may be attained with a less expenditure of time and teaching effort.

An efficient standard. The second qualification of a satisfactory standard is that it be "efficient." By this it is meant that the standard must represent a degree of ability which equips pupils for meeting present and future demands with a high degree of efficiency. The word efficiency has been borrowed or rather adopted from the field of engineering and mechanics. The efficiency of a machine such as a steam engine is the value of the fraction whose numerator is the amount of work which the engine does, or its accomplishment or output, and whose denominator is the amount of energy put into it in the form of fuel. The value of this fraction may be increased in two ways: first, if the numerator, that is, the amount of work done, is increased without increasing the amount of energy put into the engine, or at least without increasing it in the same proportion; second, if the denominator is decreased without decreasing the numerator in the same proportion. The most efficient machine is the one for which the value of this fraction is the largest.

The word efficiency with essentially this meaning is now employed with reference to many forms of human endeavor. The numerator consists of the actual accomplishment. The denominator consists of materials, energy, and time which are put into the project, both in the form of preparation and in the actual doing at the time. For example, the contractor for a building erects a tower for hoisting and distributing the concrete used in the construction. He installs a rock crusher and other machines and appliances which will form no part

of the completed building. All of these things are a part of the expense he puts into the building. In addition, he puts a large quantity of labor into it. These two items, together with the building material, are measurable in terms of dollars and cents and constitute the denominator of the fraction. The product of his endeavor, the completed building, forms the numerator of the fraction. His efficiency as a contractor is represented by the ratio of these two quantities.

He might have dispensed with the tower and have had the concrete distributed in wheelbarrows. He might even have dispensed with the rock crusher and mechanical mixer for the concrete. By so doing he would have eliminated these items of expense, but it is reasonably certain that if he had done so the total expense of constructing the building would have been considerably increased by the added labor, and this would have decreased the efficiency of the enterprise. The total expenditure of time and effort or money is the sum of the expenditures for preparatory and accessory purposes, plus the expenditures of actual operation. In a large project, if the expenditure for preparatory and accessory purposes is too small, then the operation expenses are unduly large, making the total larger than necessary. A high degree of efficiency demands that there be such an adjustment between the two as to make the sum as small as possible.

Effort to be expended on the tool subjects. The school subjects are frequently classified under two heads:— tool subjects, and content subjects. The tool subjects of the elementary school are reading, handwriting, the operations of arithmetic, spelling, and language. In the study of the content subjects, such as the problems of arithmetic, literature, geography, history, science, etc., the tool subjects are used. The situation with respect to school subjects is quite analogous to that of the illustration just cited. The tool subjects

are used in further learning in school, and in practical activities outside of school. Time and effort are required for acquiring skill in using these tools. Time and effort are also required when these tools are used. If only a small degree of skill is acquired, the time and effort required for using the tools are greatly increased. For example, time and effort are required to learn to add. By increasing the amount of time for practice, the skill of the learner can be increased and the time required to add numbers in the solving of problems and in the practical activities will be decreased. If the learner has a large amount of adding to do, it will be economy for him to spend a relatively large amount of time in practice, that is, in preparation. If he is going to have only a few occasions to add numbers, it will not be economy for him to spend a large amount of time in practice.

The situation is precisely the same as that of the contractor who was mentioned. When he is constructing a \$500,000 building, it is economy of time and effort (both of which have a money equivalent) to spend several hundred dollars and several weeks of time in preparation for the construction of the building. If, however, he were building a \$3000 residence it would be folly to spend a very large amount in preparation for the work. So if the pupils in our schools are going to have many occasions to add in their future school work and in their activities outside of school, it will be economy to spend enough time and effort to engender in them a relatively high degree of skill. If, on the other hand, these pupils are going to have only a few occasions to add, it is folly to expend the time and effort to engender in them a high degree of skill. What is true of addition is true of the other operations in arithmetic and of the skills involved in the other tool subjects.

School demands on the tool subjects. Some of the occasions for the use of these tools occur in the work of the school,

and some occur outside of school in practical activities. It is generally conceded that these tools should be acquired in the first six grades. In the seventh and eighth grades and the high school pupils have many demands made upon them for reading, writing, spelling, the operations of arithmetic, and expression by means of language, both oral and written. Our manner of carrying on school work by the use of textbooks and reference libraries makes the demand for reading very heavy. It is also our custom to require much written work, prepared outside of the recitation period, and in some subjects much written work during the recitation period. This custom makes heavy demands for writing, spelling, and written expression. In arithmetic we expect pupils to learn to solve problems (not examples) by solving problems. In fact we require them to solve many problems, and the solving of problems requires arithmetical operations to be performed. In view of the fact that the school itself makes enormous demands upon its pupils for the use of these tools, it is folly not to prepare them adequately for these demands. It would be just as sensible for a contractor of a \$500,000 building to fail to provide a mechanical mixer for concrete as for a school to fail to prepare its pupils to read with an appropriate rate and quality of comprehension. In the case of the contractor, failure to provide appropriate machinery means that the concrete must be mixed by means of back-breaking and time-consuming labor. In the case of the school, failure to equip the pupils properly to read means that the numerous assignments which they will be asked to read will not only consume an enormous amount of time, but will also destroy interest in the school work because for them reading is a slow and difficult and hence a disagreeable process.

Outside of school there are a number of demands for these tools which are common to all: — reading newspapers, mag-

azines, and books; writing letters; expressing ideas; and solving simple arithmetical problems of everyday life. In addition to these there are a number of special demands which depend upon one's occupation. Educators differ concerning the extent to which public schools should prepare pupils for these special demands, but many agree that little differentiation should be made below the seventh grade, and, therefore, the question of preparation for these special demands concerns us but little in the consideration of standards for the tool subjects.

Basis for standards of accomplishment. The demands of the school, and the common demands of life outside of school, are the requirements which are to be considered in the setting up of standards for the tool subjects in the elementary school. In general discussions concerning what the schools should accomplish, and in practically all of the discussions of particular standards, attention has been focused upon the demands of life outside of school, and the demands of the school have been overlooked. This is perhaps due to the recent emphasis upon the fact that the function of the school is to give children preparation for the activities of life outside of school. This is a most wholesome and commendable point of view, but its acceptance should not blind one to the fact that the demands which the activities of the school make for the use of these tool subjects exceed many of the demands which the common activities outside of school make. The average man or woman does not meet as pressing demands for reading as do the pupils in the high school. Likewise the demands for writing, and probably for the other tool subjects as well, which pupils meet in school are greater than they will meet outside of school.

By saying that they are greater it is meant, not only that the demands are more numerous, but also that they involve a limited time for their satisfaction. For example, when a

pupil is given an examination in school it is not only necessary that he write legibly, but it is also very necessary that he write reasonably rapidly and without focusing his attention upon the act of writing. If he does not he is seriously hindered in answering the questions.

Even if the tool subjects were not practical, it would be necessary for the school to teach them and teach them well in the first six grades, in order that the pupils might do the work of the following grades. There is no valid basis for the argument that, since few of the pupils will become bookkeepers or clerks, or enter other specialized occupations, emphasis upon definite standards of skill in performing the operations of arithmetic and in handwriting is unjustifiable.

It should, however, be recognized that in the case of the content subjects the source of the standards of attainment is the demands of life outside of school. For example, if we were considering standards for the solving of problems in arithmetic instead of the doing of examples, the source of standards would be the demands which exist in the practical activities of life.

It should be evident from the foregoing discussion that a standard may be set too high. On the other hand it may be too low. Either condition means a low degree of efficiency. A teacher should not take pride in the fact that she has brought her pupils up to a point well above the standard. This condition may mean that she is just as inefficient as the teachers whose pupils are below standard, her inefficiency being due to an unusual expenditure of time for the engendering of this particular outcome.¹

Translating the scores into school marks. The first step in translating the scores for a class into school marks is to

¹ For further consideration of norms (standards) read Monroe, Walter S., *An Introduction to the Theory of Educational Measurements*, chap. VIII (Boston: Houghton Mifflin Company, 1923.)

determine the general status of the class. After the first weeks of a term an experienced teacher will be able to form some estimate of the general status of his class. The giving of a general intelligence test will be helpful in this connection. The distribution of the I.Q.'s of a class may be considered a very reliable index of their general status. If the median I.Q. is below 100, the teacher may know that he has on the average poor pupil material. If the median I.Q. is above 100, he may know that his class consists of pupils better than the average. If there is a relatively large number of low I.Q.'s, it may be expected that there will be a relatively large number of low "grades." A standardized achievement test will also be helpful.

A few test-makers have worked out a plan for translating scores into "grades." (See Courtis Silent Reading Test No. 2, Courtis Supervisory Tests in Arithmetic and Geography,¹ and the Burgess Picture Supplement Scale for Measuring Reading Ability.) In the absence of a specific plan recommended by the test-makers the following general procedure may be used in translating test scores into school "grades." The same plan can be followed in translating point scores on written examinations into school "grades."

The scores of the class may be arranged in ascending order of magnitude; as: 32, 35, 38, 40, 41, 45, 46, 47, 50, 51, 54, 55, 56, 57, 58, 60, 63, 64, 68, 69, 70, 73, 74. There are 23 pupils in this class. The median score is 55. The translation of this median score into the corresponding "grade" is the first step. If the class is an average one, this median score of 55 should be translated into the median or average "grade" which the school recognizes. If "grades" are reported in terms of per cents and the passing mark is 75, the average "grade" will be approximately 85. Hence the score of 55

¹ The plan which Courtis gives does not include school marks, but they could be substituted for the descriptive terms which he uses.

will be translated into a "grade" of 85. If "grades" are reported in letters, such as A, B, C, D, E, the score of 55 would be equivalent to a "grade" of C. In case the class is not a typical one, the median score should not be taken as corresponding to the median "grade." For example, suppose the class is known to be a superior one. The median "grade" might be 90 or B. In extreme cases it might be even higher. If the class is made up of inferior pupils, the median "grade" will be below 85 or C.

The determination of the grade corresponding to the median score furnishes a basis for the translation of the other scores. If we take the median score of 55 as being equivalent to a grade of C, the grades corresponding to the other scores are indicated in the distribution below. If it is desired, numerical grades may be used instead of the letters A, B, C, D and E.

		58		
		57		
	47	56	69	
	46	55	68	
38	45	54	64	74
35	41	51	63	73
32	40	50	60	70
E	D	C	B	A

This distribution is for an average or typical class. If the class is superior, and the median score 55 has been translated in a "grade" of B —, the distribution of the scores might be as follows:

			64	
		54	63	
		51	60	74
	41	50	58	73
	40	47	57	70
	38	46	56	69
32	35	45	55	68
E	D	C	B	A

The general status of the class is the determining factor in translating scores into school "grades." In the case of standardized achievement tests, this will be indicated by comparing the median score of the class with the grade norm. The information yielded by a general intelligence examination will also be helpful. When using unstandardized tests or ordinary examinations, the teacher will have to rely upon his experience with other classes and information obtained from other sources in estimating the general status of the class.

Use of the normal probability curve. There is a somewhat prevalent opinion that the normal probability curve fixes the per cent of pupils who should receive "grades" below the passing mark. This is a mistaken notion. The normal probability curve tells us nothing concerning the per cent of pupils who should receive any "grade." It is true that from a statistical point of view there are certain divisions of the curve which are convenient. If the base line of the curve is limited to a length equal to five times the standard deviation (5σ) and this distance is divided into intervals of 1σ and perpendiculars are erected at the division points, the per cent of cases falling in each division of the curve will be 7, 24, 38, 24, and 7. It has been suggested that these numbers define the per cent of pupils who should receive grades of A, B, C, D, and E, respectively. Similar proposals have been made for other plans of dividing the normal probability curve.

No plan of dividing the normal probability curve can claim to be distinctly superior to any other plan. One may accept, without being inconsistent, the assumption that the "grades" for a large group of pupils should conform to the normal probability curve, and at the same time refuse to accept any particular proposed specifications as to the per cent of pupils who should receive each "grade." For example, a distribution of 50 per cent A's, 25 per cent B's, 10

per cent C's, 9 per cent D's, and 6 per cent E's is not necessarily inconsistent with the assumption that accurate measures of achievement of unselected groups tend to form a normal distribution. However, in this case, it would be necessary to define a "grade" of A as being a "grade" which means that the pupil is above average ability or, in other words, that all pupils who are average or above are given a "grade" of A without any attempt to distinguish between their achievements. Thus a "grade" of A would represent a wide range of achievement. On the other hand, "grades" of B, C, D, and E would represent narrow ranges of achievement. The different "grades" would not represent equal ranges of achievement. This, however, for certain purposes, may not be entirely undesirable.

A standard distribution regarded as an item of school policy. The range of achievement which a "grade" shall represent, or, in other words, the per cents of pupils who in the long run shall receive the different "grades," is a matter of school policy. This should be determined by the school. Perhaps the best way to define the range of achievement which a "grade" is to represent is in terms of the per cent of pupils who, in the long run, will receive the "grade." Undoubtedly, at the present time, there is considerable variation from system to system in respect to the range of achievement that is represented by the different "grades." If a system can reach a definite agreement concerning the per cent of pupils who should receive each mark, a step will be taken in the direction of making objective the norms which are used in translating examination scores into school marks. If only five marks are used a school would probably not be far from the general practice if the per cents of pupils receiving these "grades" were defined as 7, 24, 38, 24, and 7. However, in establishing this standard distribution it should be distinctly recognized that it is done as a matter of school

policy instead of being forced upon the school by the nature of the normal probability curve.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What is the distinction between "scores" and "school marks"?
2. What things must be known in order to translate scores into school marks?
3. What do you think of the plan of placing standardized test scores on pupils' report cards instead of school marks? (It is assumed that if scores are used the standards will be given also.)
4. Why are school marks, such as "92 per cent," "good," "excellent," and the like, indefinite?
5. Why are norms necessary?
6. Why are averages and medians not satisfactory norms?
7. How must satisfactory norms be determined?
8. What is the meaning of "efficiency" in education?
9. How is a pupil's score interpreted?
10. What different types of norms are available?
11. What is a "standardized" test?

CHAPTER XIII

ADMINISTRATIVE AND SUPERVISORY USES OF EDUCATIONAL TESTS

IN the preceding chapters standardized tests and scales have been considered primarily from the point of view of their usefulness to the teacher in diagnosing pupils and classes. They are also valuable to the administrator and supervisor. It is the problem of this chapter to set forth the relation of these school officials to the use of standardized tests and scales, and to consider how they may be used by them in organizing and directing the work of the school.

The supervisor's responsibility. The supervisor ¹ (superintendent, principal, or special supervisor) must assume the responsibility of educating the teachers of a school system in making and using educational measurements. Standardized educational tests are so new that few teachers have become acquainted with them in the course of their professional training. It is important that the teacher think of standardized tests and scales as instruments which will enable him to make his instructional efforts more effective. There is considerable evidence to show that if a teacher misinterprets the function of tests or looks upon them with suspicion, his efficiency as an instructor will be lowered by their use. The supervisor should train the teachers to give the tests properly, or arrange to have some trained person give them to all classes.

There should be uniformity in the tabulation of the scores. If directions and blanks for this work are not furnished

¹ In order to avoid awkward phraseology, "supervisor" will be used to refer to both administrator and supervisor.

with the tests chosen, the supervisor should decide what directions are to be followed. If the tabulation requires much time, it is probably unwise to require the teachers to do it. Substitute teachers, and in some cases normal-training students, can be utilized for this purpose if trained clerical help is not available. In addition to the tabulations which may be made by the teachers, the supervisor should make others to show the situation for the school as a whole. These are helpful not only to the supervisor, but also to the teacher. He needs to see his work in relation to that of the school as a whole. The supervisor should assume responsibility for making the teachers aware of the complete significance of the information which the tests yield. It is his function to provide norms and comparable scores from other cities, if they are not easily accessible to the teachers. The significance of a group of facts is more easily grasped when they are represented graphically. The supervisor can render valuable service to the teachers by preparing a chart or series of charts to show the standing of the several grades of the school in comparison with the norms.

The value of a test depends upon the use made of the measures. The tabulation of the scores yielded by standardized educational tests and their interpretation do not constitute use. In order to use the measures which such tests yield it is necessary to make them the basis of action. We have already indicated in connection with the description of particular tests some of the remedial instruction which may be instituted by the teacher in response to the conditions revealed by the tests. In addition to such uses, measures of achievement and general intelligence may be utilized in the administration and supervision of the school. In some of these procedures the teacher has a part, but they concern primarily the administrator and the supervisor.

Without this step standardized tests become mere "play-

things" and their use cannot be justified. The omission of this step creates a situation similar to that which would exist if a physician examined a patient carefully and determined the nature of his ailment, but did not prescribe any remedial treatment. In our zeal to convert teachers to the acceptance of the principle that the measurement of certain results of instruction is possible, there has been a tendency to overlook this step. In fact some have even said that they were content to apply the tests and reveal to the teachers the shortcomings of their work. These persons would leave to the teachers the difficult problem of remedying the defects. As a result not a few teachers have failed to see in the tests anything more than a new "plaything," which they might use to secure material for a paper to read at a teachers' association or to arouse the interest of their pupils. Such teachers have expressed their approval of the tests when their pupils' scores were high, and have considered the tests unsatisfactory when the scores were low.

Administrative and supervisory uses of standardized educational tests. The major activities in which the supervisor may use standardized educational tests are the following:

1. Promotion and classification of pupils.
2. Educational and vocational guidance.
3. Supervision of instruction, including evaluating of school efficiency.
4. Rating of teachers. (This is partly included in the above activity.)
5. School publicity.
6. Scientific experimentation.

Each of these activities ¹ depends upon other factors, but measures of achievement and general intelligence constitute

¹ Only the first two will be considered in this book. See also Monroe, Walter S., *An Introduction to the Theory of Educational Measurements*, chap. x. (Boston: Houghton Mifflin Company, 1923.)

one of the important bases for action. None of these activities are new except possibly the second and the last. The others have been an essential phase of supervision for many years. Before standardized educational tests were available, measurements of the achievements of pupils were made by cruder methods. Now that it is possible to make more accurate and more precise measures of these abilities, our attention has been directed to these activities in connection with standardized educational tests. In discussing the uses of standardized tests indicated by the first and second of the activities enumerated above, considerable space will be devoted to the activities as such. This is done because it is not possible to make intelligent use of test scores without a clear understanding of the activity to which they are applied.

Two fundamental questions which should not be confused. In considering the administrative and supervisory uses of measures yielded by standardized educational tests there are two fundamental questions which should not be confused. The *first* relates to the desirability of the administrative or supervisory procedure in connection with which measures of achievement or general intelligence may be used. The *second* question concerns the accuracy and validity of the measures yielded by the test. For example, in the case of educational guidance it is necessary to consider separately the desirability of educational guidance or of any particular method of educational guidance, and the validity and accuracy of the measures yielded by the standardized educational tests used in connection with this work. It may be highly desirable to advise students concerning the courses which they should undertake, and even to insist upon the suggestions being followed, providing accurate and valid information is available for making the recommendation. On the other hand, it may be undesirable to insist upon any student following a particular course. The wisdom of edu-

cational guidance is one question and should not be confused with the accuracy and validity of measures yielded by educational tests. It is true that the two questions are interrelated. If the measures are grossly inaccurate or lacking in validity, it would probably be unwise to use them as a basis for a procedure such as the promotion or classification of pupils.

Test scores not the only basis of action. The administrative and supervisory activities considered here should not be based solely upon measures of achievement and general intelligence. Other traits of pupils, such as initiative, effort, interest, health, etc., should receive direct consideration. This is especially true in the case of promotion and classification of pupils, and in educational and vocational guidance.

Basis of determining the value of a given procedure. That administrative or supervisory procedure is most valuable which tends to make the school most efficient. This, however, does not necessarily mean that the procedure which results in the greatest degrees of skill and information is the most valuable. Skills and knowledge represent only a portion of the outcomes of instruction. Ideals, attitudes, prejudices, and perspectives must also be considered. These outcomes are very subtle, and at the present time we are unable to measure them directly by means of standardized tests. They appear to be influenced by the methods of instruction, the organization of the school, its general spirit, and particularly the composition and spirit of the group of pupils who are brought together for instructional purposes. It is likely that certain policies of promotion or classification of pupils, when extended over a period of years, would be very potent in determining the general attitude and atmosphere of the school, which in turn would be very potent factors in engendering these subtle but important outcomes. Thus, in determining the value of a procedure it is necessary to con-

sider all outcomes engendered by the school and also to insist upon these outcomes for all children.

1. Promotion and Classification of Pupils

The organization of our school systems. The typical American school system is organized in twelve grades. In the first six of these and, in many cases, in the first eight, the pupils are promoted by grade. During the last four years and in some schools beginning with the seventh grade, promotion is by subject. Pupils enter about the age of six and are assigned to the first grade. At the end of the year those who are judged to have done satisfactory work are promoted to the second grade. This procedure is repeated each year until the pupil finishes the elementary school. If a system of semiannual promotion is followed, readjustments are made at the middle of each year.

It is a well-established fact that pupils differ widely with respect to their capacity to learn. Some learn very rapidly and very well. Others learn slowly, and a few appear to be unable to make appreciable progress in the regular work of the school. In order to adjust the school to these differences in capacity, some reclassification of pupils is made at each promotion time. Most of these adjustments consist of requiring pupils who have not made satisfactory progress to repeat the work which they have just gone over. In some school systems such pupils are promoted on trial. This plan gives all a chance to do the work of the next grade if they are able. Occasionally pupils who are exceptionally capable are permitted to skip a grade, but this occurs very rarely in most school systems. In the high school the plan of promotion is by subject rather than by grade and students are not permitted to skip a subject. Bright students are, however, occasionally permitted to carry more than the usual number of subjects. Those who are judged to do the work

unsatisfactorily receive no credit for the subject and are required to take it over or to take another in its place. Thus, the reclassifying of pupils in a typical American school consists primarily of refusing to promote those who have done unsatisfactory work.

Need for further adjustment of the school to the capacities of pupils. The application of standardized tests of both achievement and general intelligence has demonstrated that, under our present system of adjusting the school to the capacities of pupils, those who are grouped together for instructional purposes exhibit a wide range of individual differences. In a typical grade or class there will be some pupils who are able to advance much more rapidly than the class as a whole. Others are able to do the work only with much difficulty, and still others appear to be unfitted to do the tasks required of them. Many educators assert that a homogeneous group is necessary for a high degree of efficiency in instruction.¹ It is obvious that such a group cannot be secured under our present plan of classification and promotion. Thus, there is need for some additional adjustments of the school to the capacities of the pupils.

Flexible plans of promotion. Since 1890 a number of plans of promotion have been advocated which are more flexible than the general plan outlined above. In these flexible plans of promotion there has been an effort to adjust the school to the capacities of the children. In general they have provided for the rapid progress of the more capable children and the slow progress for those who are unable to maintain the regular pace. Among those plans which have attracted considerable attention are the Cambridge Dual-Track Plan, the Pueblo Individual Plan, the Portland Cycle Plan, and the Batavia Plan. Recently emphasis has been

¹ Experimental proof of this assertion is lacking, although a number of arguments can be advanced in support of it.

given to a multiple-track plan which includes more than mere flexibility of promotion. An essential feature of the plan is the establishment of differentiated courses of study for the different groups.

Terman's five-track plan. Professor L. M. Terman is one of the most ardent advocates of the multiple-track plan of classifying pupils. It is his conviction that "provision should be made for five groups of children: the very superior, the superior, the average, the inferior, and the very inferior. We may refer to these as classes for the 'gifted,' 'bright,' 'average,' 'slow,' and 'special' pupils. For each of these groups there should be a separate track and a specialized curriculum."¹ Soon after entering school, children would be assigned to the "track" appropriate to their level of intelligence. According to Professor Terman's plan each of these tracks would constitute in many respects a separate school. The course of study would be different and there would be many differences in methods of instruction. The pupils in the different tracks would also advance at different rates. Although a pupil would be assigned to a track soon after entering school, the road for transfers from track to track would be kept open so that readjustments could be made in the case of pupils who had demonstrated their ability to do the work of the next higher track, or had failed to do satisfactorily the work of the track to which they had been assigned.

Such a plan as Professor Terman proposes would be applicable only to a large city school system. In smaller systems it would be necessary to limit it to a three-track plan, and perhaps in small towns and villages it would be necessary to limit the organization to two tracks.

The establishment of a multiple-track system not dependent upon the use of educational tests. In considering the

¹ Terman, L. M., and others, *Intelligence Tests and School Reorganization*. (Yonkers: World Book Company, 1922. 19 pp.)

desirability of a multiple-track system it is necessary to keep in mind that such a system could be instituted without the use of any standardized tests of achievement or general intelligence. Pupils could be segregated on the basis of their school records and teachers' estimates, at least beyond the first grade. The limitations of educational measurements, particularly measurements of general intelligence, do not constitute an argument for or against any multiple-track plan of school organization, except as these limitations constitute obstacles to the realization of the plan.

Is a multiple-track plan of the school organization desirable? At the present time this appears to be a debatable question. Scientific data on which to base a final answer are not available. Professor Terman's proposal of a multiple-track plan of school organization has many supporters and a worthy number of opponents. The question of the desirability of such a plan of school organization is an important one. It has been asserted that the welfare of our democracy is at stake. It appears likely that the answer given to the question will have a very significant influence upon our future life as a state and as a nation. If we should perchance give the wrong answer, we may expect serious consequences to follow. Within a generation or at most within two generations the ideals and general organization of our social group can be completely changed by the education which the children receive in our public schools. We should, therefore, clearly understand the question at issue. As we have already pointed out, it should be divorced from questions relative to the validity and reliability of educational tests. It should also be understood that partial segregation of pupils for instructional purposes is now practiced in most schools, particularly in our high schools. In many communities the student has considerable option in the subjects which he shall undertake. In some high schools there are two or more dis-

tinct courses, such as commercial course, vocational course, general course, etc. Even in the elementary school, as a result of failures and extra promotions there is a gradual selection of pupils in the successive grades. The question, therefore, is not whether we shall have any differentiation or segregation of pupils, but rather how much.

In order that the reader may comprehend more clearly the issues involved and the position of the writer with reference to this important feature of school organization we present in the following pages a summary of the arguments for and against a multiple-track plan of organization such as Professor Terman proposes. The results of certain investigations will be reviewed briefly. The reliability of educational tests both of achievement and of general intelligence has been treated in the preceding chapters. Only very brief consideration of this supplementary question will be given here.

Arguments for a multiple-track system. Many of our prominent educators are advocating this plan of organizing our schools with much eloquence and enthusiasm. They point out that at the present time many children fail in their school work. In some high schools more than one fourth of the children fail in certain subjects. The achievements of many other pupils are distinctly unsatisfactory even though they are given a passing mark. There is evidence that this condition is due in part to a lack of the adaptation of the school to the capacities and interests of many pupils. It is claimed that if the course of study and methods of instruction were suited to the capacities of these children, many of them would be able to do the work with at least a fair degree of success. It is also pointed out that there are many bright children who fail to find in the work of the school, as it is now organized, a real challenge to their capacities. They are required to spend much time upon drill which they do not

need. Much of the work is so easy that it fails to be interesting to them. Difficult tasks are divided and subdivided until these bright children become bored with the anæmic character of the work they are given an opportunity to do and with the plodding rate of progress which they are forced to maintain. Not infrequently it is said of some children that they have wasted a year or even more because if they had been given an opportunity they could have advanced more rapidly. It is maintained that the achievements of both the bright and the dull would be much greater under one of the proposed multiple-track systems.

The advocates of segregation of school children on the basis of their capacity to learn urge the social importance of discovering and training the more capable children for leadership. One writer, after calling attention to the large number of failures and the lack of adequate training for children on the lower levels of intelligence, says: "But much more serious is the situation caused by our failure to select and properly educate the gifted among our young people. . . . The selection and training for leadership in a democracy is the most important function to be performed by our public education. Unless we select and train geniuses, society must slip back into barbarism."¹

Another writer has discussed this question even more frankly. He points out that the "popular idea of democracy is a delusion." Our leadership has always been invested in a few. These have been men and women possessing high degrees of intelligence. This writer goes on to speculate concerning the future:

One's intelligence quotient will eventually be known and persons will be classified thereby. Those of high intelligence will be directed into the lines of occupation which call for leadership. Those

¹ Graves, F. P., "Public Provision for the Education of Adults"; in *School and Society*, vol. 25, p. 390. (April 7, 1922.)

persons will naturally be placed in the professions, and in the leading positions in industry, commerce, and politics. Each person will then be directed on a scale of intelligence down to those whose work is of the most routine character of which the imbecile is capable. But what effect will this have on our so-called democracy? It must inevitably destroy universal adult suffrage, by cutting off at least twenty-five per cent of the adults, those whose intelligence is so low as to be incapable of comprehending the significance of the ballot. On the other hand, it will throw the burden and responsibility of government where it belongs, on those of high intelligence, and we come back to the rule of the aristocracy — this time the real and total aristocracy. For its own salvation the state must assume the obligation and the responsibility of selecting this intellectual aristocracy and having selected it see that it is properly trained.¹

Arguments against a multiple-track system. This, however, is only one side of the question. There are many equally prominent educators who deprecate the point of view of the "educational determinists," a title which has been given to those who support the plan of segregation just described. They insist that the segregation of children on the basis of their mental capacities in either the elementary school or the high school will, in the long run, be fatal to the welfare of our democracy. They point out that a serious weakness of our present social organization is the lack of common understanding between persons engaged in different vocational activities and also between persons living in different sections of the country. As an illustration they cite the industrial unrest and the resulting strikes and lockouts which we are told are manifestations of a lack of understanding and confidence between capital and labor. The increasing differentiation and specialization of our activities and institutions tend to separate rather than to promote a better common understanding between individuals belonging to different groups in our state and nation. It is impossible for

¹ Cutten, George B., "The Reconstruction of Democracy"; in *School and Society*, vol. 16, pp. 477-89. (October 28, 1922.)

them to have many contacts in their vocational activities. There is much differentiation in the means of employing their leisure time. Our newspapers, as well as political parties, are partisan. The public school is the one institution which all classes of people have in common. We must, therefore, look to the public schools to provide an education which will serve as an antidote to the increasing specialization and differentiation which we find in adult activities. It is insisted that we shall fail in this endeavor if we permit any grouping of pupils on the basis of their mental capacities. Instead we should organize our schools so that each child will be brought into intimate contact in the schoolroom with children on other levels of intelligence. Thus, by having all children pursue the same course of study with a minimum of differentiation, at least until the end of the eighth grade, we shall tend to promote a better common understanding between different adult groups. The lack of adjustment of our schools to the capacities of the children is admitted. The opponents of segregation agree that the high per cent of failures is a symptom of inefficiency, but they would remedy the existing situation by improving methods of instruction and by making adjustments within classes rather than by reorganizing the school as a whole.

Scientific determination of merits of multiple-track plan by scientific experimentation. Theoretically, it appears that one should be able to determine the merits of a multiple-track plan of school organization by means of scientific experimentation. Practically, it is exceedingly difficult or impossible to do so because many of the important outcomes are very subtle and of such a nature that we are at present unable to measure them. Furthermore, it would be necessary to continue such experimentation over a period of several years in order to observe the ultimate effect upon the attitude and atmosphere of the school.

Several investigations have been carried on for the purpose of collecting data in regard to the merits of a multiple-track plan of organization. The most pretentious of such investigations which has come to the attention of the writer was carried on in eight public schools in the city of Chicago.¹ This experiment extended over three semesters and involved an average enrollment of approximately eight thousand children. In the four experimental schools the pupils were classified into grades largely upon the basis of their mental age as determined by two general intelligence tests.² The pupils were grouped into slow, average, and fast sections within each grade largely on the basis of their intelligence quotients.

The results of this investigation are not at all conclusive. Several important factors were not considered. The measurement of achievement was confined to the more formal and tangible outcomes of instruction. Such outcomes as industry, good citizenship, attitudes toward fellow pupils, honesty, social development, etc., were either not measured in this experiment or measured so indirectly that no conclusions can be made concerning their presence or amount. If the experiment has been continued for a longer time, say eight or ten years, certain effects would likely have been noted that did not appear during the three semesters, or effects that were present might have appeared in more pronounced fashion. A slight superiority in the more formal achievements³ was shown for the four schools which were

¹ Odell, Charles W., *The Use of Intelligence Tests as a Basis of School Organization and Instruction*. University of Illinois Bulletin, vol. 20, no. 17, Bureau of Educational Research Bulletin, no. 12. (Urbana: University of Illinois, 1922. 78 pp.)

² In Grades I-B to III-B inclusive, these tests were the Pressey Primer Scale and Dearborn Group Intelligence Tests. In Grades III-A to VIII-A inclusive, they were National Intelligence Scale A and Illinois General Intelligence Scale.

³ Principally operations of arithmetic and silent reading.

organized on the basis of the three-track plan. The superiority was very slight. The investigator, after considering the limitations of the experiment, expresses the opinion that the advantages to be gained are sufficient to justify the additional expense involved. He, however, hastens to add the warning that such a multiple-track organization is not "a panacea for all inefficient schools nor a method of organization that should be rushed into by every school administrator before he has made a careful study of its installation and operation."

Statements by other investigators. A number of other investigators have made less extensive studies of the effect of a multiple-track system of classification. Although their conclusions should not be given much weight, they are significant in that they indicate popular approval of the plan. The following quotation is among the more conservative:

If group intelligence scores were to be used for classifying pupils into small groups of homogeneous ability, we could apparently expect a great many mistakes, but the real significance of this would depend upon how far a given pupil is out of place, how serious for the purpose in hand is such a displacement, and, in a practical sense, upon how much better even such a classification is than the hit-or-miss grouping which usually prevails. As a matter of fact, nothing is commoner in educational literature at present than favorable and even enthusiastic reports of experiments in classification on the basis of scores on some group intelligence test.¹ Disregarding the possibility that where the plan fails, the experiment is not written up, this would seem to show that great accuracy in ranking the pupils is not essential, at any rate for a most noticeable improvement over present practice.²

¹ Jordan, R. H., "An Example of Classification by Group Tests"; in *Educational Administration and Supervision*, vol. 6, pp. 198-201. (April, 1920.)

² Geyer, Denton L., "Reliability of Rankings by Group Intelligence Tests"; in *Journal of Educational Psychology*, vol. 13, pp. 43-49. (January, 1922.)

Miller ¹ lists the following advantages for the classification of high-school pupils on the basis of mental ability:

1. "It makes possible an adaptation of the technique of instruction to the needs of the group." Under this head the author discusses the need for allowing more time for some pupils to formulate an answer to a thought question than others. He states that it is his conviction at the present time that thought questions are put at a rate too rapid for a large majority of the class.

2. "Classification makes possible, but does not insure, an adaptation of materials of instruction to the needs of the group."

3. "Classification may make competition operative as an incentive." Miller's notion here is that the incentive of competition is stronger when it is between pupils of approximately the same capacity.

Accuracy of classification on basis of test scores. In connection with the description of the various tests we have given facts with reference to their reliability. It will, however, not be out of place to give certain additional facts relative to the accuracy of the classification which may be expected from test scores. Geyer ² gave the Otis Group Intelligence Test and the Illinois General Intelligence Scale to one hundred and twenty pupils in the junior high-school grades of the Chicago Normal School. The coefficient of correlation between the two sets of scores was .642. He, however, states:

If these one hundred and twenty pupils had been divided on the basis of the intelligence scores of one test into four class-sections of

¹ Miller, W. S., "Administrative Use of Intelligence Tests in High Schools"; in *Twenty-First Yearbook of the National Society for the Study of Education*, part II, p. 205. (Bloomington, Illinois: Public School Publishing Company, 1922.)

² Geyer, Denton L., "Reliability of Rankings by Group Intelligence Tests"; in *Journal of Educational Psychology*, vol. 13, pp. 43-49. (January, 1922.)

ordinary size, 51.6 per cent of them would have been in the wrong section according to the other test, and 31.8 per cent of them would have been out of place by an amount equal at least to half the range of such a class-section.

The Thurstone and the Brown University tests were given to fifty-four freshman college students. If these had been grouped into two classes on the basis of the scores of one test, 26 per cent of them would have been in the wrong class according to the other. For a sophomore group of sixty-four students a classification on the basis of scores of one test would have been changed in 32.8 per cent of the cases by a reclassification on the basis of the second test.

Breed and Breslich¹ have presented some illuminating data relative to the use of scores yielded by general intelligence tests for the classification of pupils. The Chicago Group Intelligence Test, Form A, the Otis Group Intelligence Test, Advanced Examination, Form A, and the Terman Group Test of Mental Ability, Form A, were given to a group of seventh-grade pupils and also to a group of ninth-grade pupils. The average intercorrelation between the scores yielded by the three tests was .77. The authors show, however, that if the pupils were classified into three sections according to scores yielded by one test, thirty per cent would be found out of place according to one of the other tests.

In the ninth grade where there were sixty pupils concerned, achievement tests in mathematics were administered monthly and a final examination was given at the end of the semester. If the classification of pupils on the basis of average achievement is compared with the classification on the basis of intelligence, it is found that out of a total of

¹ Breed, F. S., and Breslich, E. R., "Intelligence Testing and the Classification of Pupils"; in *School Review*, vol. 30, pp. 51-66, 210-26. (January, March, 1922.)

fifty-one pupils for whom complete records were obtained, twenty-eight, or fifty-five per cent, were displaced.

It should be noted that we are considering here the accuracy of classification on the basis of only intelligence test scores. If such scores are combined with other information concerning the pupils, a more accurate classification may be expected. However, we do not have data on which to base a statement of the probable accuracy of classifications made in this way.

The ultimate effect of segregation of pupils upon society. The idea of providing each child with an educational opportunity especially adapted to his capacity to learn is attractive. It appeals to our sense of fairness. However, one must not forget that the welfare of the total social group is paramount. Our schools are maintained by society in order to give children the education which is necessary for the preservation and advancement of civilization. The welfare of the group rather than the welfare of the individual is the goal to be attained. Usually, there is no conflict between what is good for the individual and what is good for society, but when there is a conflict, society must always come first. Our prisons and houses of correction are evidence of the attitude of society toward individuals who attempt to contribute to their personal welfare in a way that is detrimental to the larger group. It is probable that either plan of school organization will result in an injustice to some children. There is no doubt at the present time that adequate educational opportunities are not provided for all children who are enrolled in our schools, but it is not at all certain that under the proposed plan of segregation there would be no injustice to any child. In fact, it appears altogether likely that a few children will be wrongly classified even when precautions are taken. Thus the fact that particular cases of injustice can be cited for either plan must not be accepted as a condemnation of that plan.

It is highly important that the child's education include all elements which are essential to his effective participation in all of the activities of adult life. Achievements in school subjects are important, but they do not constitute the total necessary equipment. All of one's education does not come from books. Ideals, perspectives, and attitudes are highly important controls of conduct. It may be that an intimate association in the classroom of children on different levels of intelligence is necessary to the engendering of certain essential elements of one's education. Leaders are needed, but it does not necessarily follow that the kind of leaders which we need can be developed by segregating the brighter children and providing them with a special education. One of the qualities for successful leadership is the ability to understand other people. Leadership involves service. "It seems, indeed, to be the verdict of human experience that the less they [leaders] have their eyes on the 'distinctions' and 'prominence' that the term 'leadership' so inevitably implies, the greater will be their chance of becoming effective leaders in a democratic society. Nothing seems to be more inimical to such leadership than the overweening consciousness of one's superiority to the common run of humanity."¹ It does not appear unlikely that, if the brighter children were segregated in the school and pursued a special course of study, there would be fostered a feeling of superiority. Unconsciously this attitude might be built up by the teachers and by the community. Such a feeling of superiority would be incompatible with our democratic ideals and would mean a lowering of the efficiency of our schools.

The effect of segregation of pupils upon the school. It is also necessary to take into consideration the possible effect of the deterministic point of view upon the general attitude

¹ Bagley, W. C., "Professor Terman's Determinism"; in *Journal of Educational Research*, vol. 6, p. 383. (December, 1922.)

which teachers would maintain with reference to their work. It does not appear unlikely that they would tend to acquire a fatalistic attitude toward their pupils. They would know in advance that certain children could not be expected to do successfully certain school subjects. In the case of any given pupil they would know in advance about the quality of work which he could be expected to do. Thus it would be futile for the teacher to attempt to inspire him to greater achievements because they would be impossible for him.

One can only speculate as to the ultimate effect of the acceptance of this deterministic point of view. It is not inconceivable that, as a result, our compulsory attendance laws might come to be modified so that they would apply only to the brighter children. The cost of maintaining our schools is rapidly becoming a very heavy burden for many communities. If there are many children who are unable to take advantage of many of the educational opportunities now provided, it would not be inconsistent with the deterministic point of view to limit attendance at our schools, particularly high schools and colleges, to the children on the higher levels of intelligence. In fact, some of the opponents of this point of view have pointed out that this procedure would be only a logical result of the segregation of children on the basis of their intelligence. This plan is beginning to be followed in some of our private colleges and is definitely implied in one of the statements already quoted. (See page 441.)

The value of complete segregation problematical. The merits of an extreme plan of segregation of school children such as Terman proposes have not been determined. It is still a question which should engage the attention of educators, particularly those working in the field of educational research. The limitations of educational tests is an obstacle to the realization of the plan, but with appropriate checks these could be overcome. It is likely that the more formal

achievements of many if not all of the pupils would be increased by means of segregation, but it also appears likely that there would be losses. At the present time a superintendent should give very thoughtful consideration to the matter before reorganizing his school system on either a three-track or a five-track plan. The possible ultimate effects should be considered as well as the more immediate and more readily measurable achievements in the formal subjects of the curriculum.

Partial segregation recommended. As pointed out in the beginning the question at issue is, "What degree of segregation is desirable?" In the past there has been relatively little and our schools are not now organized so that they are well adapted to the capacities of the children. In the first four or five years of the school, we may very properly permit the more capable children to gain one or two years. In the upper grades and in the high school, they may be permitted to carry additional subjects. Furthermore, there should be systematic educational guidance in the choice of the subjects or courses which they elect. In some schools we find rapid progress sections in some subjects for bright children and slow progress sections for dull ones. Usually this grouping is accompanied by only slight changes in the subject-matter of the course. The principal difference is in rate of progress. Such differentiation appears to be beneficial and not subject to some of the disadvantages of complete segregation.

Other plans of adjusting the school to the pupil. A satisfactory adaptation of the school to the child cannot be obtained merely by allowing some to progress more rapidly and others more slowly than the normal rate. There must be some adaptation of the course of study to the different levels of the children to be instructed. When instructing groups which are not homogeneous with respect to capacity to learn

it is possible to devise a plan whereby the same assignment will not be given to all members of a class. Those who are less capable should be given an opportunity to restrict their efforts to the minimum essentials of the course and the bright children should be given assignments which will furnish a challenge to their capacities. Furthermore, each pupil should be given an opportunity to do all that he is able to do. Not only this, he should be encouraged to exercise all of his talents. It is possible to work out such a plan. In order to be successful, it would probably require a revision of the meaning of school marks. At the present time our school marks A, B, C, etc., are intended to describe the quality of the work which pupils do. Of course, as a matter of fact we know that, in addition to differing in quality, they also differ in the amount of work which they do. It is possible to divide our assignments into sections corresponding to our marking system. According to this plan a minimum assignment limited to the essentials of the subject would be required for the lowest passing mark. Additional assignments would be required for each of the higher marks. In order to attain a given mark a student would have to do satisfactorily all of the assignments required for that mark. All pupils should be encouraged to complete as many sections of the assignment as they are able. This plan has actually been followed in some schools, but it has not received the recognition to which it is entitled.

The application of this plan rests largely with the teacher. In order to be efficient in applying it, the teacher must know her pupils. Good teachers have always studied their pupils. Some are very skillful in becoming intimately acquainted with their qualities and limitations. Others are not so skillful. It has been shown in many instances that pupils have been misjudged by their teachers. Some have been rated as dull or even feeble-minded when they were exceptionally

able. General intelligence tests furnish a means of securing information about our pupils. To most teachers they will be exceedingly helpful, but it should be remembered that these tests cannot furnish all the information which a teacher needs. He should be aware of the likes and dislikes of his pupils, their attitude toward the school, the motives to which they respond — in fact, he should be in possession of all information which is useful in guiding and assisting children in their education. Only when teachers know their pupils in this way can we expect to secure an efficient adaptation of the school to the child. A satisfactory adjustment can never be secured by means of any mechanical plan of organization.

2. Educational and vocational guidance of pupils

The meaning of educational and vocational guidance. The promotion and classification of pupils which we have just considered represent one form of educational guidance. However, educational guidance includes much more. In junior high schools and to a greater extent in the senior high school some election of work is permitted. Whenever the school offers the pupil an opportunity to choose either the courses or the particular subjects which he will pursue, there is need for guidance in making this choice. In so far as this guidance relates to the child's vocation activities after leaving school it becomes vocational guidance.

Need for educational and vocational guidance. The present need for educational and vocational guidance grows out of certain conditions. Formerly a highly selective group of boys and girls attended our high schools. A very large per cent of them had definite intentions of going to college and preparing for some professional career. Thus there existed a relatively high degree of unanimity of interests both for the subjects studied in the high school and also with respect to their vocational expectations. Now the

situation is materially changed. There has been a phenomenal increase in the number of boys and girls attending high school; many of them have no intention of going on to college. Thorndike has estimated that in 1918 approximately one of every three children reaching their teens in the United States entered high school. In 1890 the corresponding figure has been estimated to be one in ten. The change which has taken place from 1890 to 1918 may be indicated by saying that "for every one hundred children who reach the age of fourteen there were approximately three and one half times as many beginning high school in 1918 as in 1890."¹

In the absence of accurate measures of general intelligence a generation ago, it is impossible to do more than estimate the probable change in mental capacity of the children who attend high school. There is, however, evidence which shows that the increase in the relative number of children who enter high school has been accompanied by a corresponding decrease in their average intelligence and an increase in the range of intelligence of each age group. Pupils differ in their interests and their ambitions. Some have acquired good habits of study, while others have not. Some invest much effort in their school work, while others are inclined to shirk assignments. Pupils also differ in personal characteristics. Some have acquired habits of fluent speech, while others express their ideas with difficulty. Some have winning personalities; others do not.

We have broadened our concept of the function of secondary education to include preparation for the various activities of adult life. As a result the number of subjects offered has been greatly increased, especially those that prepare for some vocational activity. A student cannot take all of the subjects offered. He must make a selection. One

¹ Thorndike, E. L., "Changes in the Quality of Pupils Entering High School"; in *School Review*, vol. 30, p. 367. (May, 1922.)

plan is to group the subjects in courses, such as commercial course, scientific course, general course, industrial or vocational course, etc. When a student elects a course, he thereby elects the particular subjects which he will study. Under another plan English and a few other subjects are required of all students. Additional subjects are to be elected by the student subject to certain restrictions. In some schools there is a combination of the two plans.

The need for intelligent guidance of students in the selection of the subjects which they pursue is apparent. Thorndike¹ has recently reported an analysis of the programs of school subjects actually being taken by pupils in ten school systems. The results show an astonishingly large number of different programs elected by the students within a single school system. For example, in one school system from which reports were secured for 139 tenth-grade students, there were 110 different programs reported. In another school with 60 students reporting there were 45 different programs.

The results of the investigations are summed up as follows:

The plain fact is that, except for the almost universal requirement of English for two years or more and for the very common requirement of algebra during the first year, high-school programs have very little uniformity. . . . Within a generation the high-school course has changed from an offering of certain almost prescribed programs, classical, Latin-scientific, English, commercial, and the like, to an offering by isolated subjects. Whatever arrangements restrict election of studies permit a student to take programs which are almost, if not quite, as varied as the programs of college students at Harvard during its period of substantially free election.

¹ Thorndike, E. L., and Robinson, Eleanor, "The Diversity of High-School Students' Programs"; in *Teachers College Record*, vol. 24, pp. 111-21. (March, 1923.)

The large per cent of failures in high school an index of inefficient guidance. Investigations have shown that in some high schools as many as one fourth of the children enrolled in certain subjects fail to do the work successfully. In mathematics the average per cent of failure is in excess of twenty, and it is only a little lower for Latin.¹ A number of causes tended to produce this large per cent of failures. Some are due to lack of effort on the part of the student; some are probably due to poor teaching; and others to improper home environment, irregular attendance, etc. However, it is probably true that a considerable number of those who failed were lacking in capacity to do the work, and if appropriate advice had been given them the time devoted to the subject in which they failed could have been more profitably invested.

Educational guidance should reduce but not eliminate failure. The high per cent of failures which now prevails in many high-school subjects was given as one of the reasons why some systematic plan of educational guidance should be instituted. From this it may be inferred by some that there should be no failures in a school where an efficient system of educational guidance prevails. Such an inference should not be made. Some pupils are lazy and uninterested in their work. Although it is conceivable that the lack of interest might be noticeably decreased under a system of educational guidance, it is not likely that it would be entirely eliminated. In addition there are some pupils who have very little ability. Thus, if we are to maintain defensible standards of work we may expect some failures even when a highly efficient system of educational guidance prevails, but the per cent of failures should be materially reduced.

Guidance in business. The fact that persons may be

¹ O'Brien, F. J., *High-School Failures*. Teachers College Contributions to Education, no. 102, p. 21. (New York: Teachers College, 1919.)

misfits in the vocation which they enter is emphasized by the attention which large commercial concerns have given to the selection and guidance of their employees. The employment office fulfills a guidance function. Applicants are given tests which have been found useful, and they are accepted and assigned to work largely on the basis of the results of these tests. In addition, those employees who make unsatisfactory records are frequently tested in order to ascertain the cause. There have been many cases in which an employee was found to be unsuited for the work in which he was then engaged, and by being transferred to another department he achieved marked success.¹ The following cases are illustrative:

"M" was a graduate of the University of Chicago. That she possessed a keen mind no one who talked with her five minutes could doubt. An English college woman had just achieved wonders in the cost-accounting department, and was calling for an assistant. "M" was given the job. Her first task was to learn to compute percentages on a slide rule. After several days of patient effort her superior appealed for an exchange. "M" could not do the work. Still she was recommended as unusually intelligent, willing, with pleasing personality. She was sent to be tested. The results were extremely striking. Below the lower quartile in all arithmetic tests, she was almost without equal in the excellence of all other scores. A vacancy was created for her in a correspondence group, and in a short time she was head of the group, auditing all correspondence which went out and dictating many letters herself from technical data furnished by the engineers. By the use of tests in the beginning she might have been saved the discouragement of the first failure, which might under less favorable conditions have lost a valuable employee to the company.

"R" was such a failure. He was eighteen years old, and has

¹ Carney, C. S., "Some Experiments with Mental Tests as an Aid in the Selection and Placement of Clerical Workers in a Large Factory"; in *Procedure to the Sixth Conference on Educational Measurements*; Extension Division, Indiana University Bulletin, vol. 5, no. 1, pp. 72-73. (Bloomington: University of Indiana, 1919.)

been a newsboy, telegraph messenger, delivery boy on a department store truck, "rustler" in an express office, and now finally stock-clerk in a big factory. With no definite aim, never satisfying, holding each job a few months, then hunting another, anything so long as it paid enough money, he was on the high road to becoming a chronic failure. Nothing had ever really aroused his interest. Fortunately, instead of being "fired" he was sent to be tested. The scores were interesting. Ability was there without doubt. But in what useful line? Terman-Binet tests showed a good normal adult, but in arithmetic he was low. (He left school because he could not master decimal fractions.) He thought he would like to be an electrician, so a job was found for him in a motor repair shop. Here he was fortunate in finding a sympathetic teacher who encouraged the boy to do his best. Soon he was given outside jobs. He showed a liking for climbing, so crane and elevator jobs were passed to him. Now he is an experienced motor maintenance man, with real money in his pay-envelope, and two years of steady service safely behind him. He has found his job.

"G" was another failure. He had been tried on various tasks, without success. Tests showed more than normal ability. Nineteen years old, he was still drawing the pay of an errand-boy. He had been a beginner for five years, and had "gone stale." It was decided to arouse his ambition by giving him a big promotion. He was made assistant to the clerk who ordered raw material from the blue-prints, at nearly twice his former wage. He entered evening classes in drafting and shop practice, and applied himself to his work with a will. In six months his whole personality altered. He spoke with assurance, held his head up, walked as though he were going somewhere, and had in every way "made good."

Educational guidance policies. There seem to be three fairly distinct policies of educational guidance: (1) Where the policy is one of "enlightenment," the emphasis is placed upon making the pupils acquainted with the opportunities and requirements of the vocation which they may enter. There is also an effort to enlighten them in regard to the subjects which they may select. This type of educational guidance does not really require individual conferences with pupils. (2) A policy of "monitory" guidance requires in-

dividual conferences with pupils. In addition to enlightening them in regard to school subjects and vocational activities, they are warned concerning their probable lack of success if they make certain choices. Monitory guidance would involve the use of educational tests, particularly those of general intelligence. (3) "Pigeon-hole" guidance is the application of the thesis that pupils should be assigned to certain courses or subjects on the basis of information secured by means of educational tests and through other sources. Under this system of guidance the choice is not left to the pupil. His educational future is determined for him by the school. This is an application of the deterministic point of view to educational guidance.

These policies overlap, and in actual practice two or more of them may be combined. Thus the school might carry on a campaign of "enlightenment" to supplement a "monitory" or "pigeon-hole" system of educational guidance.

Information needed in educational and vocational guidance. 1. *Pupils' capacity to learn.* One of the most important items of information concerning the pupil is an accurate knowledge of his capacity to learn. His previous school record is one index of this capacity. Another index can be secured by means of general intelligence tests such as were described in Chapter IX. There are a few prognostic tests which have been designed to measure a pupil's capacity to succeed in a particular subject. Examples of this type of test are the Wilkins Prognosis Test in Modern Languages, the Rogers Test of Mathematical Ability, and the Van Wagenen Reading Scales in History, English, and General Science.

2. *Pupils' interests.* In advising pupils it is necessary to be acquainted with their interests. Other things being equal, children will do much better in a subject in which

they are interested than in one in which they are not. Occasionally a pupil of mediocre ability will succeed in a school subject or a vocation because of his intense interest.

3. *Vocational information.* There is also need for information concerning the opportunities or needs for various types of training. It would be a serious mistake to advise pupils to pursue a course of study which would result in preparation for a given vocation unless it is probable that there would be an opportunity to enter that vocation. In some particular vocations there is little demand and unwise guidance might very easily create a surplus of applicants.

4. *Relation of degrees of capacity to probable success.* In interpreting information in regard to a pupil's capacity to learn, it is necessary to know the relation between various degrees of capacity, and probable success in the pursuit of certain school subjects and in vocational activities.

Relation between general intelligence and school success. The relation between degrees of capacity and school success has been determined by calculating the degree of correlation between scores on intelligence tests and school grades. The coefficients of correlation for different groups of pupils exhibit a considerable range. Trabue¹ gives the following coefficients of correlation between measures yielded by intelligence tests and average scholarship marks:

Otis Group Intelligence Tests (C.B.).....	.535
Otis Group Intelligence Tests (Scores).....	.470
Mentimeter Test.....	.481
National Intelligence Test, Form A.....	.459

These coefficients are based upon one hundred and twenty pupils and the scholarship record is for one semester only.

¹ Trabue, M. R., "The Influence of Intelligence Tests in Junior High Schools"; in *Twenty-First Yearbook of the National Society for the Study of Education*, part II, p. 186. (Bloomington, Illinois: The Public School Publishing Company, 1922.)

Miller¹ reports a coefficient of .522 between scores yielded by the Miller Mental Ability Test and scholarship. Proctor,² using the Army Alpha Intelligence Test, secured coefficients of correlation with school marks of .343 and .413. Colvin³ reports the following correlation between scores yielded by the Otis Group Intelligence Test and school marks:

Forty-four correlations calculated between the test scores in the Brookline schools and the scholarship records give the lowest Pearson coefficient 0.40 and the highest 0.91. The median is 0.69. Seven of these coefficients are between 0.40 and 0.49, eight between 0.50 and 0.59, eight between 0.60 and 0.69, thirteen between 0.70 and 0.79, seven between 0.80 and 0.89, and one between 0.90 and 1.00.

In Miss Wheeler's School, correlations between scores in the tests and school standing were computed for the five upper grades. These varied from 0.253 and 0.801 when calculated grade for grade. The correlation of the scores of all the pupils tested in these five grades with their school marks was 0.487.

For a group of 124 eighth-grade pupils the correlation between C.B.'s and average marks in four school subjects was found to be .586.

Relation between general intelligence and occupational success. The most extensive study of the relation between levels of intelligence and occupational success is given in the *Report of Psychological Examining in the United States*

¹ Miller, W. S., "Administrative Use of Intelligence Tests in High School"; in *Twenty-First Yearbook of the National Society for the Study of Education*, part II, p. 215. (Bloomington, Illinois: The Public School Publishing Company, 1922.)

² Proctor, W. M., *Psychological Tests and Guidance of High-School Pupils*. Journal of Educational Research Monograph no. 1, pp. 15, 16. (Bloomington, Illinois: Public School Publishing Company, 1921.)

³ Colvin, Stephen S., "Recent Results Obtained from the Otis Group Intelligence Scale"; in *Journal of Educational Research*, vol. 3, pp. 1-12. (January, 1921.)

Army.¹ From a study of this report and other similar investigations it appears that success in a given occupation is not likely by one whose general intelligence is below a certain level. However, there are a number of occupations corresponding to most levels of intelligence. We cannot, therefore, say that, given a pupil's level of intelligence, we can advise him explicitly with reference to the particular occupation he should enter. We can only say that his success is unlikely in those occupations for which the average level of intelligence is materially above his intelligence, and that his intelligence is likely to be wasted in those occupations for which the average level of intelligence is below his own.

Difficulties encountered in educational and vocational guidance. There are certain limitations of educational and vocational guidance which should not be overlooked. In the first place, it is difficult or impossible to secure all of the evidence which is needed. In carrying on a guidance program it will not be possible to collect exact information concerning a large number of personal traits of pupils which have a marked significance with respect to their future success. Furthermore, the evidence which we are able to collect is not perfectly accurate and we do not have accurate norms for interpreting much of the information which we can secure. In predicting the future success of a pupil either in school or in a vocation, we are dealing only with probabilities and not certainties. A forecast of the future is only a prediction. This prediction is based upon certain assumptions in regard to the stability of the capacities to learn, interests, and other traits of pupils. Often under an efficient system of educational guidance some predictions

¹ Yerkes, Robert M. (edited by), *Psychological Examining in the United States Army*. National Academy of Sciences Memoir, no. 15, pp. 819 ff. (Washington: Government Printing Office, 1921.)

will not be realized. However, the evidence available at the present time clearly indicates that the number of misfits can be materially reduced by advising pupils in regard to the subjects and courses which they should pursue.

Administration of educational guidance. A prerequisite to the successful administration of a plan of educational guidance is an adequate system of records. An individual record blank should be devised which has space for an accumulative record of the pupil's school career. The more important items of information are the following: name of pupil, chronological age, nationality, economic status of parents, health, previous school history, scores on standardized tests, educational plans, and vocational ambition. If this information is gathered at the time the pupil enters high school, additional entries of test scores and school grades should be made as they are available. In the case of high schools which draw pupils from a number of elementary school buildings, it will be helpful to have the information indicated above gathered and placed in the hands of the high-school principal or some one designated by him prior to the admission of the student into the high school. A number of high-school principals have found it helpful to encourage the giving of a general intelligence test and certain achievement tests to the pupils completing the eighth grade in their district.

The collection of information in regard to students is only the first step. If effective guidance is to result, it must be used intelligently. In order to insure this there should be a definite organization for this purpose and some one person be made responsible for the work. As we have already indicated, teachers need to know their pupils. For this reason as well as others it seems desirable to have the teachers participate in the educational guidance of students. However, they should be assisted and directed by some one who has

made a special study of the problems involved. In most high schools this person should be the principal. When the demands of other administrative matters are too great to permit the principal to give sufficient time to this work, it may be delegated to a vice-principal or some other person who is qualified for the work. In large city school systems where there are several high schools, it is desirable to have a department of educational guidance attached to the office of the city superintendent. The particular plan of organization will depend somewhat upon the general administrative organization of the school system.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What is the deterministic point of view?
2. What do you understand by "segregation of pupils?"
3. What evidence would be necessary to prove that a homogeneous grouping of pupils is essential to the highest degrees of teaching efficiency?
4. Distinguish between "educational guidance" and "vocational guidance."
5. Would you approve of a plan of promoting all children on trial? Give your reasons.
6. Should we expect that some one will soon determine scientifically the merits of Terman's proposal for segregation? Give your reasons.

SELECTED BIBLIOGRAPHY

- Bagley, W. C. "Vocational Guidance and the Teacher of Science"; in *School Science and Mathematics*, vol. 13, pp. 89-97. (February, 1913.)
- Bagley, William C. "Educational Determinism; or Democracy and the I.Q."; in *School and Society*, vol. 15, pp. 373-84. (April 8, 1922.) Also in *Educational Administration and Supervision*, vol. 8, pp. 257-72. (May, 1922.)
- Bagley, William C. "Professor Terman's Determinism: A Rejoinder"; in *Journal of Educational Research*, vol. 6, pp. 371-85. (December, 1922.)
- Bennett, H. S., and Jones, B. R. "Leadership in Relation to Intelligence"; in *School Review*, vol. 31, pp. 125-28. (February, 1923.)
- Berry, Charles S. "The Classification by Tests of Intelligence of Ten Thousand First-Grade Pupils"; in *Journal of Educational Research*, vol. 6, pp. 185-203. (October, 1922.)
- Branson, Ernest P. "An Experiment in Arranging High-School Sections

- on the Basis of General Ability"; in *Journal of Educational Research*, vol. 3, pp. 53-55. (January, 1921.)
- Breed, F. S., and Breslich, E. R. "Intelligence Tests and the Classification of Pupils"; in *School Review*, vol. 30, pp. 210-26. (March, 1922.)
- Brewer, John M. "Guidance in the High School with Special Reference to College Entrance"; in *School Review*, vol. 29, pp. 434-43. (June, 1921.)
- Clerk, Frederick E. "The Arlington Plan of Grouping Pupils according to Ability in the Arlington High School, Arlington, Massachusetts"; in *School Review*, vol. 25, pp. 26-47. (January, 1917.)
- Cleveland, Elizabeth. "Some Further Studies of Gifted Children"; in *Journal of Educational Research*, vol. 4, pp. 195-99. (October, 1921.)
- Counts, George S. "Education for Efficiency"; in *School Review*, vol. 30, pp. 493-513. (September, 1922.)
- Counts, George S. "The Population of the Private Secondary Schools"; in *School and Society*, vol. 15, pp. 570-73. (May 27, 1922.)
- Counts, George S. "The Selective Principle in American Secondary Education, II"; in *School Review*, vol. 30, pp. 95-109. (February, 1922.)
- Cowdery, K. M. "A Statistical Study of Intelligence as a Factor in Vocational Success"; in *Journal of Delinquency*, vol. 4, p. 227. (November, 1919.)
- Dickson, Virgil E. "Use of Group Mental Tests in Guidance of Eighth-Grade and High-School Pupils"; in *Journal of Educational Research*, vol. 2, pp. 601-10. (October, 1920.)
- Dickson, Virgil E. "What First-Grade Children can do in School as Related to what is shown by Mental Tests"; in *Journal of Educational Research*, vol. 2, pp. 475-80. (January, 1920.)
- Dvorak, August. "Recognition of Individual Differences in the Junior High School"; in *School Review*, vol. 30, pp. 679-85. (November, 1922.)
- Edgerton, A. H. "Present Status of Guidance Activities in Junior High Schools"; in *Education*, vol. 43, pp. 173-83. (November, 1922.)
- Flersheim, E. "Duties of the Vocational Counselor in the High School"; in *Chicago Schools Journal*, vol. 4, p. 88. (November, 1921.)
- Freeman, Frank N. "Bases on which Students can be Classified"; in *School Review*, vol. 29, pp. 734-45. (December, 1921.)
- Fretwell, Elbert K. *A Study in Educational Prognosis*. Teachers College Contributions to Education no. 99. (New York: Teachers College, Columbia University, 1919. 35 pp.)
- Geyer, Denton L. "The Reliability of Rankings by Group Intelligence Tests"; in *Journal of Educational Psychology*, vol. 12, pp. 43-49. (January, 1922.)
- Glass, James M. "Classification of Pupils in Ability Groups"; in *School Review*, vol. 28, pp. 495-508. (September, 1920.)
- Holmes, Henry W. "The General Philosophy of Grading and Promotion in Relation to Intelligence Testing"; in *School and Society*, vol. 15, pp. 457-61. (April 29, 1922.)

- Hughes, W. H. "Provisions for Individual Differences in High-School Organization and Administration"; in *Journal of Educational Research*, vol. 5, pp. 62-71. (January, 1922.)
- Kelley, Truman Lee. *Educational Guidance: An Experimental Study in the Analysis and Prediction of Ability of High-School Pupils*. Teachers College Contributions to Education no. 71. (New York: Teachers College, Columbia University, 1914.)
- Ketner, Sarah P. "Grouping by Standardized Tests for Instructional Purposes"; in *Journal of Educational Research*, vol. 2, pp. 620-25. (October, 1920.)
- Kitson, H. D. "Psychological Tests and Vocational Guidance"; in *School Review*, vol. 24, pp. 207-14. (March, 1916.)
- Leavitt, Frank M. "School Phases of Vocational Guidance"; in *School Review*, vol. 23, pp. 687-96. (1915.)
- Lemon, Harvey B. "Forecasting Failures in College Classes"; in *School Review*, vol. 30, pp. 382-87. (May, 1922.)
- Madsen, I. N. "Group Intelligence Tests as a Means of Prognosis in High School"; in *Journal of Educational Research*, vol. 3, pp. 43-52. (January, 1921.)
- Madsen, I. N. "Intelligence and Success in High School"; in *Journal of Educational Research*, vol. 3, pp. 396-98. (May, 1921.)
- Madsen, I. N. "The Contribution of Intelligence Tests to Educational Guidance in High School"; in *School Review*, vol. 30, pp. 692-701. (November, 1922.)
- Maveriak, Lewis Adams. "The Class in Occupations"; in *School and Society*, vol. 16, pp. 348-51. (September 23, 1922.)
- Mead, A. D. "Orientation Course for Freshmen at Brown University"; in *School and Society*, vol. 3, p. 428. (March 18, 1916.)
- Miller, W. S. "The Administrative Use of Intelligence Tests in the High School"; in *Twenty-First Yearbook of the National Society for the Study of Education*, part II, pp. 189-222. (Bloomington, Illinois: Public School Publishing Company, 1922.)
- O'Brien, Francis P. *The High-School Failures*. Teachers College Contributions to Education no. 102. (New York: Teachers College, Columbia University, 1919. 97 pp.)
- Odell, Charles W. *The Use of Intelligence Tests as a Basis of School Organization and Instruction*. University of Illinois Bulletin, vol. 20, no. 17, Bureau of Educational Research Bulletin, no. 12. (Urbana: University of Illinois, 1922. 78 pp.)
- Pintner, Rudolf, and Noble, Helen. "The Classification of School Children According to Mental Age"; in *Journal of Educational Research*, vol. 2, pp. 713-28. (November, 1920.)
- Proctor, W. M. "The Use of Intelligence Tests in the Educational Guidance of High-School Pupils"; in *School and Society*, vol. 8, pp. 473-78, 502-09. (October 19 and 26, 1918.)

- Proctor, W. M. "The Use of Psychological Tests in the Educational Guidance of High-School Pupils"; in *Journal of Educational Research*, vol. 1, pp. 369-81. (May, 1920.)
- Proctor, W. M. "Psychological Tests as a Means of Measuring the Probable School Success of High-School Pupils"; in *Journal of Educational Research*, vol. 1, pp. 258-70. (April, 1920.)
- Proctor, W. M. "The Use of Psychological Tests in the Vocational Guidance of High-School Pupils"; in *Journal of Educational Research*, vol. 2, pp. 533-46. (September, 1920.)
- Proctor, W. M. *Psychological Tests and Guidance of High-School Pupils*; in *Journal of Educational Research*, Monograph no. 1. (Bloomington, Illinois: Public School Publishing Company, 1921.)
- Rugg, H. O. "Rating Scales for Pupils' Dynamic Qualities: Standardizing Methods of Judging Human Character"; in *School Review*, vol. 28, pp. 337-49. (May, 1920.)
- Ryan, W. Carson, Jr. *Vocational Guidance and the Public School*, U. S. Bureau of Educational Bulletin, no. 24. (Washington, 1918.)
- Seashore, C. E. "Sectioning Classes on the Basis of Ability"; in *School and Society*, vol. 15, pp. 353-58. (April 1, 1922.)
- Terman, Lewis M. *The Intelligence of School Children*. (Boston: Houghton Mifflin Company, 1919. 317 pp.)
- Terman, Lewis M. "The Use of Intelligence Tests in the Grading of School Children"; in *Journal of Educational Research*, vol. 1, pp. 20-32. (January, 1920.)
- Terman, Lewis M. "The Psychological Determinist; or Democracy and the I.Q."; in *Journal of Educational Research*, vol. 6, pp. 57-62. (June, 1922.)
- Terman, L. M., Dickson, V. E., Sutherland, A. H., Franzen, R. H., and Fernald, G. *Intelligence Tests and School Reorganization*. Subcommittee Report, N.E.A. (Yonkers: World Book Company, 1922.)
- Theisen, W. W. "Provisions for Individual Differences in the Teaching of Reading"; in *Journal of Educational Research*, vol. 2, pp. 560-71. (September, 1920.)
- Thorndike, E. L. "Changes in the Quality of the Pupils Entering High School"; in *School Review*, vol. 30, pp. 355-59. (May, 1922.)
- Thorndike, E. L., and Symonds, P. M. "The Occupations of High-School Graduates and Non-Graduates"; in *School Review*, vol. 30, pp. 443-51. (June, 1922.)
- Trabue, M. R. "The Use of Intelligence Tests in Junior High Schools"; in *Twenty-First Yearbook of the National Society for the Study of Education*, part II, pp. 169-88. (Bloomington, Illinois: Public School Publishing Company, 1922.)
- Varner, G. F. "Can Teachers Select Bright and Dull Pupils?" in *Journal of Educational Research*, vol. 6, pp. 126-32. (September, 1922.)
- Washburne, Carleton W. "Educational Measurement as a Key to Individ-

ual Instruction and Promotions"; in *Journal of Educational Research*, vol. 5, pp. 195-206. (March, 1922.)

Whitney, Frank P. "Provision for Accelerant and Retarded Children in Junior High School"; in *School Review*, vol. 27, pp. 695-705. (November, 1919.)

Willett, G. W. "A Suggestion for Meeting Individual Differences"; in *School Review*, vol. 28, pp. 576-84. (October, 1920.)

Willing, Matthew H. "The Encouragement of Individual Instruction by Means of Standardized Tests"; in *Journal of Educational Research*, vol. 1, pp. 193-98. (March, 1920.)

CHAPTER XIV

IMPROVEMENT OF WRITTEN EXAMINATIONS

IN Chapter I attention was called to certain of the imperfections of written examinations as instruments for measuring the achievements of school children. It was pointed out that more satisfactory measures could be secured by using standardized educational tests and also by the improvement of the written examinations prepared by the teacher. The important standardized achievement tests have been described in Chapters II to VIII. It is the purpose of this chapter to give additional information in regard to the reliability of written examinations, and also to present in an organized way suggestions for their improvement.

Reliability of written examinations. As we pointed out in Chapter I, examination "grades" are subject to two types of errors, constant and variable. The former is illustrated in the tendency of some teachers to give high "grades" and of others to give low "grades." The presence of a variable error is indicated when different teachers assign different "grades" to the same examination paper. (See page 7.) Both types of errors are found in measures yielded by standardized educational tests.¹ The usual method of ascertaining the magnitude of the variable errors of test scores is to have the test given twice to the same group of pupils. The coefficient of correlation between these two sets of scores is an index of the magnitude of the variable errors. (See page 72.) This method is not the same as that used by Starch

¹ Monroe, Walter S., *The Constant and Variable Errors of Educational Measurements*. University of Illinois Bulletin, vol. 21, no. 10, Bureau of Educational Research Bulletin, no. 15. (Urbana: University of Illinois, 1923.)

and Elliott and others in the study of the accuracy of examination "grades." (See page 5.) In fact their investigations have been confined to the subjectivity of the marking of examination papers. In order to secure information concerning the reliability of written examinations which might be compared with that for standardized educational tests, an investigation was carried on under the direction of the writer.¹ Two examinations were given to the same pupils. In most cases the questions were prepared and the papers marked by different teachers. The coefficient of correlation between the two sets of "grades"² was taken as the coefficient of reliability of the examinations. It may be thought of as an index of the variable error of measurement.

The coefficients of reliability obtained for written examinations are assembled in Table XXXII. The highest is .95 and two of them are negative. The median coefficient of reliability, .65, may be used as a general index of the variable errors of the measures yielded by written examinations. Since essentially the same method was employed in this investigation as has been used in studying the reliability of standardized educational tests, comparisons may be made with the coefficients of reliability for the various tests described in the preceding chapters. Information relative to the reliability of a number of standardized tests is summarized here for the convenience of the reader.

The coefficients of reliability of standardized educational tests. McCall³ has stated that the "range of self-correla-

¹ Monroe, Walter S., and Souders, Lloyd B., *The Present Status of Written Examinations and their Improvement*. University of Illinois Bulletin, vol. 21, no. 13, Bureau of Educational Research Bulletin, no. 17. (Urbana: University of Illinois, 1923.)

² These "grades" were really point scores.

³ McCall, W. A., *How to Measure in Education*, p. 396. (New York: The Macmillan Company, 1922.)

TABLE XXXII. SUMMARY DISTRIBUTION OF COEFFICIENTS OF RELIABILITY FOR WRITTEN EXAMINATIONS

SIZE OF COEFFICIENT OF CORRELATION	FREQUENCY
.95	1
.90	2
.85	4
.80	4
.75	9
.70	4
.65	9
.60	8
.55	4
.50	4
.45	5
.40	2
.35	1
.30	4
.25	1
.20	0
.15	1
.10	0
.05	1
.00	0
-.05	0
-.10	0
-.15	1
-.20	1
Total	66
Median	.65

tion for many standardized tests is about .5 to about .9." The writer's experience has indicated a somewhat greater range. In Table XXXIII the reliability coefficients of a number of standardized educational tests are given. Those

TABLE XXXIII. RELIABILITY COEFFICIENTS OF STANDARDIZED EDUCATIONAL TESTS

TEST	COEFFICIENT
Illinois Intelligence General Intelligence Scale ¹	.92
Courtis Standard Research Tests, Series B ³	.87
Brown Silent Reading Test — Rate.....	.86
Courtis Silent Reading Test No. 2 — Rate.....	.85
Otis Group Intelligence Scale ⁴	.84
Monroe Standardized Silent Reading Test Revised ¹ — Rate.....	.84
Courtis Silent Reading Test No. 2 — Comprehension — No. Quest.....	.80
Starch Silent Reading Test — Comprehension — Words....	.77
Monroe General Survey Scale in Arithmetic ¹	.76
Monroe Standardized Silent Reading Test Revised ¹ — Comprehension.....	.76
Monroe Standardized Silent Reading Test Revised ¹ — Rate.....	.75
Monroe Standardized Silent Reading Test Revised ¹ — Comprehension.....	.72
Starch Silent Reading Test — Comprehension — Ideas....	.72
Indiana Attainment Scale No. 1. ³	.66
Starch Silent Reading Test — Rate.....	.62
Pressey Primer Scale ²	.59
Courtis Silent Reading Test No. 2 — Comprehension — Index.....	.58
Pressey First Grade Vocabulary Scale ³	.37
Brown Silent Reading Test — Comprehension — Quantity..	.36
Pressey Primer Scale ²	.33
Brown Silent Reading Test — Comprehension — Quality..	.19

¹ Monroe, Walter S., *The Illinois Examination*, p. 47. University of Illinois Bulletin, vol. 19, no. 9, Bureau of Educational Research Bulletin, no. 6. (Urbana: University of Illinois, 1921.)

² Pressey, L. W., "A Group Scale of Intelligence for Use in the First Three Grades: Its Validity and Reliability"; in *Journal of Educational Research*, vol. 1, pp. 285-94. (April, 1920.)

³ Unpublished data of the Bureau of Educational Research, University of Illinois.

⁴ Colvin, S. S., "Some Recent Results Obtained from the Otis Group Intelligence Scale"; in *Journal of Educational Research*, vol. 3, pp. 1-12. (January, 1921.)

for the silent reading tests by Brown, Starch, and Courtis are taken from a recent bulletin ¹ by the writer. The range in this table is from .19 to .92.

¹ Monroe, Walter S., *A Critical Study of Certain Silent Reading Tests*, pp. 33, 34. University of Illinois Bulletin, vol. 19, no. 22, Bureau of Educational Research Bulletin, no. 8. (Urbana: University of Illinois, 1922.)

In certain unpublished studies the writer has accumulated the following information: The Courtis Standard Research Test, Series B, Forms 1 and 2, were given to pupils as follows: Grade V, 89; Grade VI, 81; Grade VII, 52; and Grade VIII, 38. The thirty-two coefficients of reliability ranged from .409 to .904 with the median at .665. Forms 1 and 3 were given to a slightly larger group in each of the four grades. The thirty-two coefficients of correlation between the two sets of scores for this administration of Series B ranged from .528 to .963 with the median at .704. The Woody Arithmetic Scales, Series A, were given to several groups of pupils. Two scores were secured by using alternate items of each of the scales and the coefficient of reliability was computed by applying Brown's formula.¹ The twelve coefficients of reliability computed in this way ranged from .46 to .91 with the average at .66. Forms 1 and 2 of Monroe's Standardized Reasoning Test in Arithmetic were given to pupils as follows: fifth grade, 36; sixth grade, 92; seventh grade, 76; eighth grade, 81. The coefficients of reliability for correct principle were as follows: .530, .630, .645, and .723. For correct answer they were .518, .528, .576, and .707. Using Brown's formula the coefficients of reliability for Gray's Silent Reading Tests were computed for thirty grade groups. These coefficients ranged from .55 to .85 with the median at .67. The number of pupils per group was less than one hundred in only five cases. For several grade groups reliability coefficients were secured for Monroe's Standardized Silent Reading Tests which ranged from .222 to .907 with an average of .669.

Haggerty has computed the reliability for both Sigma 1 and Sigma 3 of his Reading Examination by having the same test repeated. In the case of Sigma 1 the interval between

¹ $r_{12} = \frac{2r_h}{1 + r_h}$ In this formula r_h is the correlation between two scores obtained from the alternate exercises of this scale.

the two applications of the test was six weeks. For two hundred children in Grades I to III inclusive the coefficient of reliability .84 was obtained. In the case of Sigma 3 the interval between the two applications was only two days. For 126 pupils from Grades V to VIII, inclusive, the coefficient of reliability was found to be .885.¹ For the sentence test alone the reliability coefficient was .769 and for the paragraph test, .806. For Thorndike's Scale Alpha for the understanding of sentences, McCall has reported a coefficient of reliability of .37. This was obtained by using a test similar to Alpha, but not considered a duplicate form. Gates² obtained reliability coefficients for the Thorndike-McCall Reading Scale which ranged from .25 to .72. All of these were for pupils belonging to a single grade. For the Burgess Picture Supplement Scale the author has given coefficients of reliability ranging from .62 to .99 for grade groups from the second to sixth grades inclusive. In each case the number of pupils was relatively small. Gates obtained coefficients of .62, .59, and .66 for three grade groups.

For the Otis Self-Administering Test of Mental Ability the author has reported an average reliability coefficient of .921 for the higher examination and of .948 for the intermediate examination. Presumably these coefficients are based on the scores secured from pupils for a sequence of several grades. For the separate tests of the Stanford Achievement Test the authors have reported coefficients of reliability based upon separate grade groups which ranged from .75 to .96. When the composite score of all the tests was used the reliability coefficient was reported as .98.

¹ Since these coefficients are based upon scores from a sequence of several grades they tend to be higher than those given above.

² Gates, Arthur I., "An Experimental Statistical Study of Reading Tests"; in *Journal of Educational Psychology*, vol. 12, p. 379. (October, 1921.)

The relative reliability of written examinations and standardized educational tests. It is, of course, obvious to any one who has had experience with either written examinations or standardized educational tests that neither type of instrument yields absolutely accurate measures of achievement. The question in which we are interested here pertains to the relative accuracy of the measurements yielded by the two types of instruments. McCall has stated that reliability coefficients for teachers' examinations are much lower than those for standardized educational tests. The data which have just been submitted indicate that the difference between the reliability of the two types of instruments is not so great as McCall's statement would lead us to expect. The median of the reliability coefficients for written examinations given in Table XXXII is .65. There are many reliability coefficients for standardized tests in Table XXXIII which are less than this. Furthermore, the citations of other coefficients of correlation in the preceding pages indicate that for a number of standardized educational tests which have been very widely used, the median of the reliability coefficients for grade groups is in the neighborhood of .65. Although some of our more elaborate standardized tests, such as the Stanford Achievement Test, the Illinois General Intelligence Scale, and the Otis Self-Administering Test of Mental Ability, may be expected to yield measures whose reliability is greatly in excess of that of typical written examinations, the conclusion seems justified that many widely used standardized educational tests yield measures which possess about the same degree of reliability as the "grades" obtained from typical written examinations prepared by teachers and other school officials.

The absolute reliability of examination grades. The statement that the reliability of a typical examination approximates that of many standardized tests and is only

slightly less than that of a number of others still leaves a doubt with reference to the absolute reliability. For practical purposes a reliability coefficient of .65 should be interpreted in terms of the variable errors of measurement to be expected. The correlation tables for eight groups having a reliability coefficient of approximately .65 were taken and the scores translated into a five-point system of school grades. It is assumed that these classes were typical and the highest scores were translated into a mark of "A," the lowest into a mark of "E." This was done in an arbitrary way, but the results indicate roughly one meaning which may be attached to a reliability coefficient of .65. It was found that in forty per cent of the cases the students received the same grades on the two examinations. In another forty-two per cent the grades which they received on the first examination were only one point higher or lower than those received on the second. For example, if a student in this group received a "D" on one examination, he received an "E" or "C" on the other. The two grades received by the remaining eighteen per cent differed by two points or more.

The constant errors in examination scores. In this investigation of the reliability of written examinations, scores and not "grades" or school marks were used. The translation of the scores into "grades" (see page 425) would probably have changed very slightly if at all the coefficients of reliability, but any constant error would probably be modified by this process of translation. When the two sets of examination scores for the same group of pupils were compared, it was found that frequently the average score on one examination was much higher than the average score on the other. This was due to the fact that one examination was much harder or the marking more severe than the other. For three of the groups the difference between the averages

of the two sets of scores was zero; for eight other groups it was one. The median difference was 6.2. In one extreme case the difference between the average scores was 50. These differences should not be taken as typical because in several cases it was evident that one examination was intended to be more difficult than the others. It should be noted that the differences between two sets of examination "grades" are not constant errors. They are merely indicative of the presence of constant errors.

Relative magnitude of constant errors in examination grades in standardized test scores. In another place¹ the writer has discussed the magnitude of the constant errors in educational tests. In cases where there has been coaching for tests, intentional or not, or disregard of standard directions, large constant errors have been introduced. In one extreme instance a constant error of over three and a half years occurred in the mental age scores of a group of children. In general, however, because of the standard directions for administering the tests and scoring the papers, the objectivity of the marking, and the norms for interpreting test scores, the constant errors in standardized tests are very much smaller, and are likely always to be smaller than those found in examinations given by teachers. However, some reduction in the magnitude of the constant errors in examination scores will result when the use of either very easy or very difficult sets of questions is avoided and when a conservative plan of marking is followed.

Explanation of the apparent contradiction between the results of this investigation and the previous studies of examination grades. The conclusions with reference to the re-

¹ Monroe, Walter S., *The Constant and Variable Errors of Educational Measurements*. University of Illinois Bulletin, vol. 19, no. 10, Bureau of Educational Research Bulletin, no. 15. (Urbana: University of Illinois, 1923.)

liability of written examinations based on this investigation apparently contradict those which have grown out of investigations of the Starch-Elliott type. The question of "Why this difference?" naturally is raised. Previous investigations have been so numerous and so uniform in the character of their results that one would be inclined probably to accept them in preference to the apparently contradictory results of the present investigation. However, a careful analysis of the procedures will reveal that the results are not necessarily contradictory. The method followed by Starch and Elliott combines both constant errors and variable errors. The "grades" assigned to the examination paper in geometry were influenced both by the subjectivity of the marking and by the tendency of some teachers to grade high and of others to grade low. The present investigation has separated the variable errors from the constant errors. We have also shown that the examination scores have in some cases involved relatively large constant errors. The extreme differences between the grades assigned to the same paper reported by Starch and Elliott (see page 5) are easily explained when it is understood that they represent the combination of variable errors and constant errors. Especially is this true when we realize that the constant errors would likely be larger for teachers of different schools, as in their investigation, than for teachers in the same school, as in the present investigation.

Improvement of written examinations. Written examinations can be materially improved by reducing the constant and variable errors of measurement. There are, however, certain other important improvements which should be effected. The elimination of errors will not insure that the examination measures the proper achievements of the pupils. As we have pointed out in the preceding chapters, a rate test does not measure the same kind of achievement that is meas-

ured by a power test. Similarly an examination might be highly reliable, but fail to measure the achievements which are most important.

Importance of the problem of measurement. A clear understanding of the problem of measurement in a school subject is a prerequisite for the preparation of a satisfactory examination in that field. As has been indicated in our discussion of the problem of measurement for the various school subjects, this involves an understanding of the nature of the abilities which are engendered, and the recognition of the educational objectives which the school is endeavoring to realize. In some subjects skills or specific habits are very prominent among the abilities engendered. In such cases the rate of work is highly important. This is true in the operations of arithmetic, silent reading, handwriting, the operations of algebra, typewriting, etc. In other subjects the emphasis is placed upon information, and in others upon the ability to use this information in solving problems or answering thought questions. In measuring such achievements rate is much less important.

A teacher's particular purpose should also influence the structure of the examinations. If he wishes to obtain diagnostic measures, the examination should be modeled after some of the diagnostic tests which we have described in Chapters II to VIII. If a general measure is desired, the questions should be representative of the portion of the subject for which a general measure is desired.

Methods for increasing the accuracy of written examinations. Suggestions for reducing the variable errors of measurement will be presented under four heads: (1) directions to students; (2) types of questions; (3) decreasing the constant errors of examination "grades"; and (4) increasing the objectivity of the marking of examination papers.

1. Directions to students. An essential feature of a

standardized educational test is the detailed directions to students in regard to the test. If the exercises which they are asked to do are at all unusual, they are explained, usually by giving one or two samples. They are also told how long they will have for the test and just what they are to do. It has been found that detailed directions of this sort tend to reduce the variable errors of measurement. Similar directions will doubtless tend to reduce the variable errors of measurement in the case of written examinations. Pupils should be told whether to answer the questions in order or to answer the easiest ones first. They should be given directions concerning the length of answers required. This is especially needed in the case of general questions, such as "discuss" or "explain." The directions may properly include admonitions with reference to spelling, arrangement of work, neatness, etc. In the case of thought questions the students should be reminded that their answers should reflect careful thinking.

2. Types of questions. One source of the variable errors in examination "grades" is the subjectivity of the marking of the papers. When a question permits of only one correct answer the marking becomes highly objective. There is no opportunity for the exercising of judgment. In spelling, a pupil's performance is either right or wrong, and our practice is to allow no credit for a performance which is not entirely correct. Thus, different teachers should assign the same "grade" to a given examination paper in spelling. A similar situation exists in the operations of arithmetic, provided a common practice is followed with reference to the amount of credit to be given for examples partly right. When a teacher is asked to exercise judgment with respect to the credit which is to be given for the answers, a variable error is introduced. Different teachers will not agree on the credit which they assign to the various questions of a given

paper. In order to overcome the subjectivity in the marking of examination papers in which the pupil has been asked to "discuss," "compare," "explain," "tell why," etc., it has been proposed that we measure the pupil's acquaintance with principles and ideas by means of certain types of exercises which permit of only one answer. The true-false exercise is one of this type which has been used extensively. Instead of asking the pupil to formulate an answer in response to a question, he is asked to tell whether a given statement is true or false. For example, instead of asking the pupil, "Why did the Puritans come to America in the seventeenth century?" we may ask him to tell whether the following statement is true or false: "Puritans came to America in the seventeenth century seeking wealth." The answer to such an exercise may be given by writing a plus sign if the pupil considers the statement true or a minus sign if he considers it false.

Another type of exercise which has been used is one in which the pupil is asked to choose from a number of proposed answers the one which is correct for the question asked. This is similar to the type of exercise found in the Monroe Standardized Silent Reading Test, Revised, the Thorndike Test for Word Knowledge, and a number of others. The answer which the pupil considers correct may be indicated by underlining it or marking it in some other way. If only one of the proposed answers is correct, the marking of an examination paper becomes highly objective. Such questions have been called "recognition exercises."

"Completion exercises" have also been used. In these the pupil is asked to fill in the words which have been omitted from statements. For example:

Revenue for paying the debts of the States after the Revolutionary War was provided by the and by due largely to influence.

The marking of an examination paper consisting of completion exercises is not highly objective unless there is only one correct word for each blank.

An examination of the descriptions of educational tests given in Chapters II to VIII will suggest other types of exercises which teachers will find helpful in reducing the variable errors in the written examinations which they prepare. As we have pointed out in several places, test-makers have exercised much ingenuity in devising exercises which will be satisfactory for testing purposes. Not all types of exercises can be utilized by teachers, but where a mimeograph is available there is a large variety of types from which a teacher may choose. Examinations consisting of exercises of the types just described are called "new examinations."

Advantages of the "new examination." Exercises of the true-false or recognition type require practically no writing on the part of the pupil. Other types require very little. Thus an examination can be made more comprehensive. It is traditional for examinations to consist of ten questions. Frequently pupils cannot write upon a larger number in the time allowed. An examination of fifty true-false exercises can be answered in twenty-five minutes or less. In the case of high-school students "new examinations" consisting of one hundred recognition exercises have been completed within twenty-five minutes. There is also a very great saving of time in the marking of the papers.

Limitations of the "new examination." It does not appear likely that the "new examination," consisting of the types of exercises we have described, will entirely replace the traditional type of examination. The "new examination" cannot be used in mathematics, except to a limited extent. It cannot be used at all in English composition. The following questions taken from Hahn's Scale for Measuring the Ability of Children in History appear to require

mental processes distinctly different from those the "new examination" calls for:

State points of similarity between the position of the United States in 1812 and her position in 1912.

Arrange the following events in order of cause and effect: Force Bill, the Carpetbaggers, Fifteenth Amendment, Negro Rule in Some of the Southern States, Ku Klux Klan.

Name the Presidents of the United States since 1892.

Furthermore, it is likely that pupils would miss valuable experience and training if they were not asked at times to compare, explain, discuss, or define. This is also true of questions in which they are asked to summarize material presented on a topic, or to apply certain principles that have been presented. Hence it is difficult to conceive of the "new examination" being a complete substitute for the traditional examination.

3. Decreasing the constant errors of examination grades. The magnitude of the constant errors in examination "grades" is determined largely by the difficulty of the set of questions and the severity of the marking of the papers. Thus, some reduction in their magnitude will result when the use of either very easy sets of questions or very difficult sets is avoided, and when a conservative plan of marking is followed. However, teachers are not able to judge accurately the difficulty of a given set of questions for a given group of students. Hence it is desirable that some other means be used in order to reduce the constant errors in examination scores.

For several years a number of educators have been urging that teachers make the distribution of their "grades" conform to a standard shape. Usually they have advised that some form of the normal distribution be adopted. This proposal has been criticized. As in any controversy there have been extremists on both sides. Among the advocates

of the use of a standard distribution are those who insist that the normal distribution explicitly tells the teacher the per cent of students who must receive "A's," the per cent who must receive "B's," etc. Cases have been reported of instructors who frankly admitted that they realized a certain student deserved to receive an "A," but that they had used up all the "A's" which the distribution allowed, and that therefore the student must be satisfied with a "B." These advocates have also insisted that the normal distribution told us that a certain per cent of students must fail in the long run. The opponents of this procedure have insisted that there was no reason why any student should fail. The quality of the work done should determine a student's "grade." Furthermore, it has been pointed out that in any group of students which happened to be brought together for instructional purposes it was extremely unlikely that the distribution of "grades" should approximate at all closely any standard distribution.

The establishment of a standard distribution of "grades" or a series of standard distributions of "grades" for a school system should not be interpreted to mean that the distribution accepted as standard is to be mechanically applied by any instructor. A standard distribution is merely a device which teachers may use in order to reduce to a minimum the constant errors in their "grades," but if it is to be helpful it must be used intelligently. Just what intelligent use means is difficult to describe. The principal function of a standard distribution is to serve as a check upon the instructor. He should keep a cumulative distribution of the grades which he assigns in each course. Whenever that for a particular class departs in any conspicuous fashion from the standard distribution, he should seek the cause. If he can find a satisfactory explanation in the quality of the work done by the various members of the class or in the gen-

eral make-up of the class there will be no need to make any changes in the "grades" assigned. On the other hand, if no satisfactory explanation can be found for his departures from the standard distribution, the instructor has secured evidence of the presence of a constant error in the "grades" of some or all of the members of the class. Any tendency to grade high or low may be detected in this way.¹

4. Increasing the objectivity of the marking of examination papers by means of uniform rules. Lack of objectivity is lack of uniformity. Any procedure which will cause teachers to agree more closely in the marking of the same paper will tend to reduce the variable errors of measurement. One procedure is the formulation of definite rules which are to be followed by all teachers. These rules should include an agreement in regard to the effect of poor writing, poor spelling, and poor English upon the student's grade as well as an agreement in regard to giving credit for correct principle, partial credit for exercises partly right or partly completed. If such rules were formulated for a school system the accuracy of the examination "grades" would be materially increased. Certain additional advantages would be derived from the general adoption by teachers of a given set of rules. If this were done a comparison of the grades given to pupils in one school with those given to the pupils in another school would have more significance than at present.

Suggested rules for the administration of examinations and the marking of the papers:

1. Make the examination relatively difficult; that is, difficult enough so that there will be few perfect scores.
2. Make the examination long enough so that every member of the class is kept busy during the entire period.

¹ The procedure to be followed in translating point scores into school marks has been described on pages 425 to 429.

3. The content of the examination should agree as closely as possible with recognized educational objectives.
4. When possible give the pupils a typewritten or mimeographed list of the questions.
5. In general questions asking for discussion or explanation indicate the completeness of the discussion or the degree of elaborateness expected in the answer.
6. State the questions so that all pupils will interpret them alike.
7. Make some use of true-false exercises and other types of questions which facilitate objective scoring. (See page 481.)
8. After the questions have been distributed (or written on the board) it is well to read them aloud to the students before they begin writing. This provides an opportunity for any student to ask about obscure points and to make any corrections which are necessary.
9. Write out, at least in an abbreviated form, the answers to all questions before beginning to mark the papers.
10. Recognize the difference between point scores and "grades." Decide upon the credit to be given for each question and mark the papers in terms of point scores.
11. Except in English do not intentionally lower a student's "grade" because of poor English, poor spelling, or poor handwriting. If it is desired to recognize these characteristics of examination papers, the student may be given a separate "grade" for them or may be refused a grade in the subject unless handwriting, spelling, and English are acceptable.
12. Read all of the answers to one question before taking up the next one. This procedure will add materially to the accuracy of the "grades."
13. Use the sorting method in grading. See page 175 for a description of the method. It is most useful in grading short papers or reports which cannot be divided into parts.
14. After the papers have been marked translate the point scores into "grades" using some standard distribution.
15. Keep a cumulative record of the "grades" which you give in each subject and compare this distribution with the normal distribution, or with the standard distribution adopted by the school. In case your distributions show a marked departure from this standard inquire concerning the cause and make appropriate modifications in your marking of the papers or in translating the point scores into examination grades.

APPENDIX A

GLOSSARY OF TECHNICAL TERMS

For the convenience of students we have assembled here definitions of the more important technical terms. In a number of instances more complete definitions, together with illustrations, are given in the body of the text. For several of these, specific page references are given.

Accomplishment quotient. (See Quotient scores.)

Accomplishment ratio. (See Quotient scores.)

Accuracy. (See Quality.)

Achievement age. A pupil's age score on an achievement test is frequently referred to as his "achievement age." It is simply the age which he has attained in his achievement. The field of this achievement may be limited to a particular subject in which case a pupil's achievement age is sometimes called his "subject age" to indicate the fact that the measure refers only to his achievement in a particular school subject. In this connection "educational age" has been used to denote the average of a pupil's achievements in a group of subjects which may be considered representative of his school progress. (See Age score.)

Achievement quotient. (See Quotient scores.)

Achievement test. Achievement is used to designate those abilities which a pupil has acquired primarily from the instruction of the school. Thus, we speak of achievement in arithmetic, in spelling, in silent reading, etc. An achievement test is one which measures these abilities. The purpose of the qualifying word "achievement" is to distinguish these tests from those which are designed to measure intelligence or capacity to achieve.

Age norms. For calculating age norms the pupils are grouped according to age. Both chronological age and mental age have been used for this purpose. Theoretically, we should obtain the same numerical results for both groupings when unselected groups of children are used since the average mental age of a chronological age group is numerically identical with the average chronological age. Unless it is otherwise stated an age norm is the median or average of scores made by pupils of the designated age. Thus the age norm for nine years is for children whose age is nine years. In calculating this norm one would probably use the scores of pupils whose ages were between eight years, six months and nine years, six months.

Age score. Age norms are used as a basis for translating point scores into age scores. For example, if the age norm for eleven years is 43, a

pupil who makes a point score of 43 is said to have an age score of eleven years. Thus a pupil's age score is always interpreted as meaning that his score on the test is equivalent to the norm for the age designated by the age score. (See Age norms.)

Attainment age. (Same as Achievement age.)

Average. The average of several quantities is their sum divided by their number. When we are dealing with relatively few quantities this definition furnishes us a statement of the procedure for calculating the average. When we are dealing with a large number of quantities and they are grouped in a frequency distribution, the short method of calculation greatly reduces the labor required. However, the average has the same meaning as when calculated by the original method.¹

Coefficient of correlation. The coefficient of correlation is a statistical device used to express a summary of the relationship which exists between two sets of facts that are paired together. Perfect correlation which is represented by a coefficient of 1.00 means that the two sets of facts are related so that the largest in one set is paired with the largest in the other, the next largest are also paired together, and so on for all pairs. Perfect negative or inverse correlation is represented by a coefficient of -1.00 which means that the largest quantity in one set is paired with the smallest in the other, the next to the largest in the first set is paired with the next to the smallest in the second, and so on. A coefficient of correlation may have any value between these two extremes. A coefficient of correlation of zero means that no relationship exists between the two sets of facts.

Coefficient of reliability. The coefficient of reliability is simply the coefficient of correlation between two sets of scores secured from two applications of the same test or from duplicate forms of it. These two applications should be separated by a relatively short time interval. For most of our educational tests the coefficients of reliability, when based upon the scores made by pupils belonging to the same school grade, range from .60 to .90. For a few tests coefficients of reliability .95 or higher have been reported.

Combined dimensions. Instead of describing each characteristic of a pupil's performance separately, the directions for scoring some test papers provide for combining the descriptions of two, or, in a few cases, three, of the dimensions in a single score. For example, when the number of exercises done correctly is taken as the pupil's score on a uniform test, we have a combination of rate and accuracy. If a scaled test is timed and the number of exercises done correctly is taken as the pupil's score we have a combination of rate, quality and difficulty. (See Dimensions.)

¹ For a description of the "short method," see Monroe, Walter S., *An Introduction to the Theory of Educational Measurements*, p. 308. (Boston: Houghton Mifflin Company, 1923.)

Completion test. A completion test consists of exercises in which the pupil is asked to supply the words which have been omitted from certain statements. (See page 481.)

Composite score. A composite score is the average of the scores yielded by several tests after they have been expressed in terms of a common unit and from a common zero point. If the scores are averaged before this reduction is made, the resulting combination will frequently be lacking in meaning because different units and different zero points are used by the different tests.

Constant error. A constant error is one which is the same for all members of a given group. This group may be a single class, a school or a group of schools. On the other hand, it may be only a division of a class as, for example, a constant error might affect only the boys in a class. A constant error may be either positive or negative, the only essential characteristic being that it is the same for all members of the group concerned. (See pages 476 and 477.)

There are two kinds of constant errors — absolute and relative. An absolute constant error has the same magnitude for all members of the group regardless of the magnitude of their scores. A relative constant error maintains a constant ratio to the magnitude of the measure. Such an error would occur in measuring a linear distance of several yards if the yard stick used was half an inch too short.

Control of testing conditions. Testing conditions include all factors other than a pupil's ability which affect or determine his performance. The most important of these factors are the following: the explanation of the tests to the pupil, the time allowed for his work, the form in which the test is presented, the pupil's physical condition, his emotional status, and the effort which he makes. Testing conditions are said to be controlled when standard testing conditions prevail. These standard testing conditions are defined in part by the directions for giving the test. (See page 100.) If the resulting scores are to be compared with the norms for a test, it is imperative that the testing conditions secured should be those for which the norms are stated.

Criterion measure. A criterion measure is any measure which may be used as a basis for comparison in order to determine the reliability and validity of the scores yielded by a given test. Teacher's estimates of a pupil's achievement, his school grade, and the composite scores from a number of tests are among the criterion measures that have been used. Occasionally, the scores yielded by one test have been used as a criterion measure for judging the reliability or validity of a new test.

Cycle test. In cycle tests the exercises vary in difficulty but they are so arranged that the variations occur in cycles. For example, in a cycle test the 1st, 5th, 9th, 13th, etc., exercises might be equivalent in difficulty. The 2d, 6th, 10th, 14th, etc., exercises would also be equivalent in difficulty. A similar condition would exist for the 3d, 7th, 11th,

and 15th exercises, and for the 4th, 8th, 12th, and 16th exercises. However, the consecutive exercises might vary widely in difficulty. A cycle of difficulty would be formed by each group of four exercises. A cycle test is useful when it is desirable to include within a single test exercises on several levels of difficulty. When such a test includes several cycles it is possible to treat it as a uniform test both in its administration and its scoring without introducing a serious error.

Derived score. Except by chance, no two tests yield point scores expressed in terms of the same unit or from the same zero point. Several proposals have been made for the calculation of a derived score which describes a pupil's performance in terms of a unit that is constant for all tests or at least for large groups of tests. Usually a point score is first obtained and this is translated into the derived score. (See Age score, Percentile score, and Quotient score.)

Diagnosis. The word "diagnosis" is used with a variety of meanings. Perhaps the meaning which is associated with it most frequently is to measure separately each of the specific abilities in a given field. (See the Monroe Diagnostic Test in Arithmetic, p. 41; Charters Diagnostic Language and Grammar Tests, p. 245; Freeman's Chart for Diagnosing Faults in Handwriting, p. 170; and Rugg and Clark Standardized First-Year Algebra Tests, p. 306.)

A semi-diagnosis is secured when, instead of measuring each specific ability separately, measures of similar abilities are combined in a single score. Thus a semi-diagnosis can be secured by such tests as the Curtis Standard Research Tests in Arithmetic, Series B, p. 27; The Illinois Standardized Algebra Tests, p. 305; and the Haggerty Reading Examination, Sigma 3, p. 118.

In analytical diagnosis a pupil's performance on a test is analyzed and particular errors which he has made are studied. Such a diagnosis can be secured from any test but the completeness of the diagnosis depends upon the comprehensiveness of the test.

A diagnosis of a class is secured when the pupils are separated on the basis of their ability. Any test which is reliable and valid yields diagnostic measures of this type.

Diagnostic test. A diagnostic test is one which yields detailed information concerning a pupil's achievement in one or more relatively narrow fields. Frequently this type of measuring instrument consists of a number of sub-tests which yield separate measures of the pupil's achievement for a variety of fields. Such a diagnostic test can be transformed into a survey test by devising some procedure for combining the scores yielded by the separate sub-tests.

Difficulty. Difficulty has been defined as that characteristic of an exercise which when present in a large degree causes a large per cent of incorrect responses and when present in a small degree is accompanied by a small per cent of incorrect responses. In other words the degree

of difficulty of an exercise is determined by the per cent of incorrect responses obtained when it is given to a large number of pupils. If certain assumptions are made concerning the distribution of the ability of the group of pupils to whom an exercise is given and the point of zero difficulty is located, the degree of difficulty of the exercise can be expressed in terms of a measure of the variability of this distribution of ability. This unit is the difference in difficulty between two exercises which are answered correctly by a certain per cent of a given group of pupils. The median deviation (P.E.) is frequently used as a unit. It is defined as the difference in difficulty between an exercise which is answered correctly by fifty per cent of the pupils and an exercise which is answered correctly by only twenty-five per cent of the same pupils. The standard deviation (S.D. or σ) is also used as a unit. It is the difference between an exercise answered correctly by fifty per cent of the pupils and an exercise answered correctly by only 15.87 per cent of the same pupils. Thus we may describe the difficulty of exercises as being 2.7 P.E., 6.3 P.E., 5.2 σ , etc.

Difficulty score. A difficulty score is a statement of the highest level of difficulty on which a pupil has done the exercises with a specified or standard degree of accuracy. This score is yielded only by scaled tests.

Dimensions of a pupil's performance. A pupil's performance is described in terms of its distinguishing characteristics. These are: (1) its amount or when produced under timed conditions, the rate of work; (2) the quality or accuracy of the performance; and (3) the level of difficulty upon which it was given. These three characteristics are sometimes spoken of as the dimensions of the pupil's performance. (See Rate score, Quality, Difficulty, and Combined dimensions.)

Discrimination. A test is said to be lacking in discrimination when it fails to give different scores to pupils who are known to differ in ability. This may happen to only a few of the pupils to whom the test is given. For example, a very easy test lacks discrimination for those pupils who make perfect scores. A very hard test is lacking in discrimination for those who make zero scores. A lack of discrimination may be indicated by other evidence. If a distribution of scores differs conspicuously from the normal distribution, when we have reason to believe that the distribution of true scores would approximate the normal, we have evidence of lack of discrimination for certain pupils. If two groups are known to differ in ability, as for example, a fifth-grade group and a sixth-grade group, a test which fails to yield a higher average score for the sixth-grade group than for the fifth-grade group is lacking in discrimination. There will also be a lack of discrimination for certain pupils if the unit used is so large that pupils who differ in ability receive identical scores.

Educational age. Educational age may be taken as the average of a pupil's achievement ages in the various school subjects or at least those which are considered most important. (See Achievement age.)

Educational guidance. Educational guidance relates to the advice to students in regard to the courses or subjects which they should undertake. (See page 453.)

Educational quotient. A pupil's educational quotient is his achievement age divided by his chronological age. (See Quotient score.)

Educational objectives, agreement with. In selecting exercises for the final form of a test they may be examined with reference to their agreement with certain educational objectives. For example, in constructing his spelling scale Ayres selected certain words on the basis of their frequency of use in adult writing. Charters selected exercises for his language and grammar tests which are in agreement with the language errors made by children. In the case of other tests the consensus of opinion of competent persons has been used as a guide in the selection of exercises. (See also Statistical selection.)

Errors. An error in a measurement is the difference between the true measure and the measure which is obtained by the instrument used. This difference may be either positive or negative. (See Constant error and Variable error.)

Exercises. The exercise is a structural unit of a test. Some of the simpler types call for a word to be spelled, an example to be worked, or a question to be answered. Other exercises are more complex. Some are large, in that they consist of several items and require much time for completion. A test usually consists of a considerable number of exercises, but occasionally of a single long exercise. (For typical exercises see descriptions of standardized tests in Chapters II to IX.)

Fore exercise. A fore exercise is a preliminary test which has for its purpose acquainting a pupil with the character of the exercises which he is asked to do in the test. The pupil's performance on the fore exercise is not included in computing his score.

Form. The term "form" is practically always used in the sense of a duplicate form. Thus a test is said to have more than one form when there are duplicate measuring instruments consisting of similar but not of identical exercises. Such duplicate forms are intended to yield equivalent measures. Hence, when the two forms are administered under exactly the same conditions, a pupil should make the same score on one form that he makes on another. Investigation has shown that, in general, duplicate forms do not yield equivalent measures even when a great deal of care has been exercised in their construction. Hence, when making comparisons between scores yielded by duplicate forms, it is necessary to know concerning their degree of equivalence and to make corrections for any differences which may have been ascertained.

The "form" of a test should be distinguished from "part" and "division." In a few cases "part" has been used with a meaning very similar to "exercise" but it is generally used to designate a section or division of the measuring instrument which is designed for certain

grades. This use is illustrated by Part 1 and Part 2 of Thorndike's Scale for the Understanding of Sentences. "Division" usually has the same meaning. In a few cases "part" has been used with a different meaning. In some cases a test has been divided into "parts" without the term being used. For example, Monroe's Standardized Silent Reading Tests consist of three parts or divisions, although neither of these terms has been used in connection with its title. Test I is designed for Grades 3, 4, and 5, Test II for Grades 6, 7, and 8, and Test III for the high school. When a measuring instrument has parts or divisions (not sub-tests), the total instrument more properly would be described as a series or a group of instruments with different parts or divisions which are designed to measure the ability of pupils on different levels.

Function. The function of a test is a statement of the ability which it is designed to measure plus a statement of the type of information which it will yield concerning this ability. A pupil's performance is completely described in terms of three dimensions. The score which a given test yields may be restricted to a single dimension or it may involve two or even three, separately or in combination. A statement of the function of the test should also include some specification of its scope. A test may be very general in scope, in which case it is called a general or survey test. If it yields measures for relatively narrow fields it is called a detailed or diagnostic test. Certain tests have a prognostic function.

Grade. The word "grade" is used to designate the terms and symbols used to express a pupil's standing on a written examination or other phases of his school work. Some of the words and symbols most frequently used are "fair," "good," "excellent," "superior," "average," "80 per cent," "95 per cent," "A," "B," "C," etc. (See Teachers' marks.)

Grade norms. Grade norms are the averages or medians of the scores made by pupils in the respective school grades. In some cases a grade refers to an entire year's work. In other cases it represents only a semester's work. Usually when grade norms are stated it is understood that there are eight years in the elementary school and four years in a high school. When such norms are applied to a system which has seven or nine years below the high school, it is necessary to make adjustments.

Index of brightness. Index of brightness has very much the same meaning as intelligence quotient. There is a difference for pupils on the higher levels of intelligence.

Index of reliability. The index of reliability differs from the coefficient of reliability in that it is the coefficient of correlation between a set of obtained scores and the corresponding set of true scores rather than the coefficient of correlation between two sets of obtained scores.

It is calculated from the coefficient of reliability by the following formula in which r_{12} represents the coefficient of reliability and r_{1t} the index of reliability.

$$r_{1t} = \sqrt{r_{12}}$$

Intelligence quotient. The intelligence quotient (I.Q.) is the quotient formed by dividing a pupil's mental age by his chronological age. It is an index of his level of intelligence. (See page 340.)

Irregular test. An irregular test is one in which the exercises vary in difficulty and are not arranged in order of ascending or descending difficulty. Irregular tests usually result when exercises are selected on some basis other than that of difficulty. When extreme irregularities are avoided irregular tests may be treated as uniform tests without introducing serious errors.

Median. The median of a set of scores, arranged in ascending or descending order of magnitude, is the middle score, or when there is no middle score it is the average of the two middlemost scores.

Mental age. A pupil's age score on an intelligence test is called his mental age.

New examination. The name "new examination" has been given to lists of exercises which have been constructed so that there is only one correct answer to each exercise. Usually these examinations require very little writing on the part of the student. (See page 482. See also Completion test, True-false test, and Recognition exercise.)

Normal distribution. A normal distribution is represented graphically by a bell shaped curve. It is symmetrical. At either extreme there are very few measures. Most of the measures are grouped near the center and there is a rather gradual decrease down to zero at the extremes. Distributions which approximate a true normal distribution are frequently spoken of as normal distributions.

Norms. The norms for an educational test are determined by having the test given to a large number of pupils belonging to several groups and by taking the average or median of these scores. Thus our present norms are the average or median achievements of pupils. In most of our uses of norms we have assumed that the average or median of present achievement is that which the pupils should achieve. It has been suggested that "standard" be used to designate the scores which pupils *should make* thereby making a distinction between "norm" and "standard," but our common practice is to use the two terms with the same meaning. A test for which norms have been determined is said to be standardized. Norms may be obtained for both grade groups and age groups. (See Age norms and Grade norms.)

Norms of attainment. These represent the final objectives to be attained. (See page 188.)

Objective. A measuring instrument is said to be objective when dif-

ferent persons using it to measure the same thing secure approximately the same result. The opposite of objective is subjective. Both of these terms are relative. No educational tests are absolutely objective but those which are rather highly objective are commonly spoken of as objective tests. The scoring of a test is said to be objective when different scorers will in general assign the same scores to the same papers. (See Subjective.)

Overlapping. The term "overlapping" is used to describe the relative position of two distributions. Its most frequent use is in the case of distributions for successive grade groups or successive age groups. The per cent of one distribution which is beyond the median or average of the other distribution may be taken as the measure of the overlapping.

Percentile scores. A percentile score describes the pupil's place in the distribution of the scores of the group to which he belongs. Consider, for example, the distribution of scores of a large number of fifth-grade pupils. Locate a pupil's score on the base line of the distribution. The position of this point can be described by telling the per cent of the total scores in the distribution which are below his score. For example, if 82 per cent of the scores are below his, he may be said to have an 82 percentile score. If a standard distribution has been secured, tables may be prepared by means of which it is relatively easy to translate any point score into the corresponding percentile score.

Performance. A pupil's performance is what he does. The performance is usually written, and for testing purposes must be such that it can be easily observed by any competent observer. A performance is sometimes described as objective, which means that the result, when observed by different persons, is the same.

Point score. A point score is the score which is yielded directly by the test. Exercises done correctly, the number of exercises attempted, and the level of difficulty reached are point scores. The magnitude of a point score depends upon the size of the unit which is usually determined by the exercises, and the length of the test. It is only by chance that two tests yield point scores in terms of the same unit and expressed from the same zero point. (See Derived score.)

Power test. The term "power test" is frequently used to describe a scaled test; i.e., one which consists of exercises arranged in ascending order of difficulty. Such a measuring instrument has been called a power test since it measures the power or ability of the pupils to do increasingly difficult exercises of the same kind. With only a slight change in the meaning, other types of tests could be called power tests when only the accuracy or quality score is used. A power test is not timed.

Practice effect. Practice effect refers to the average increase of the scores of one trial over those yielded by a preceding trial, when there

has been no opportunity for coaching between the two administrations of the test. Because of becoming acquainted with the nature of the exercises, pupils tend to make higher scores on the second trial of a test than they did on the first. This practice effect constitutes a constant error when the same norms are used to interpret the scores from both trials. The magnitude of this error varies with different tests, but in general second-trial scores are on the average ten per cent greater than first-trial scores.

Practice tests. A practice test is primarily a device for giving drill or practice. It is only incidentally an instrument for measurement. (See Curtis Standard Practice Tests in Arithmetic, page 87.)

Preliminary test. (Same as Fore exercise.)

Probable error. The probable error is a device used to describe the variability or spread of the normal distribution about its average. It is similar in this respect to the standard deviation (σ) and may be calculated from it by multiplying by .6745. If we form an interval extending from -1.0 P.E. (i.e., 1.0 less than the average) to $+1.0 \text{ P.E.}$, there will be included exactly 50 per cent of the total distribution. This makes it possible to say that the chances are just even or one to one that any item selected from the distribution at random will be included within the limits of this interval. For this reason the probable error (P.E.) is more easily interpreted as a unit for measuring the spread of a distribution than the standard deviation (σ).

Probable error of estimate. The probable error of estimate is a statistical device derived from the coefficient of correlation which is helpful in interpreting cases of "high" correlation. It may be defined as the measure of departure from the perfect correlation. This is given in terms of the median deviation or P.E. of the distribution of all the departures from perfect correlation in the pairs of scores from which the coefficient of correlation was calculated. It is calculated from the coefficient of correlation by the following formula in which P.E._{Est} designates the probable error of estimate, σ_2 is the standard deviation of the distribution of scores obtained from the second application of the test, and r_{12} is the coefficient of correlation between two sets of obtained scores.

$$\text{P.E.}_{Est} = .6745 \sigma_2 \sqrt{1 - r_{12}^2}$$

A probable error of estimate of 3.4 means that in fifty per cent of the pairs of scores there is a departure of the second score from a perfect correlation with the first score of more than 3.4 in fifty per cent of the pairs.

Probable error of measurement. The probable error of measurement bears the same relation to the probable error of estimate that the index of reliability bears to the coefficient of reliability. In other words, it is a measure of the departure of a given set of obtained scores

from perfect correlation with the corresponding true scores. It is calculated from the coefficient of reliability by the following formula in which P.E. M is the probable error of measurement, σ is the average of σ_1 and σ_2 and r_{12} is the coefficient of correlation between two sets of obtained scores.

$$P.E.M = .6745 \sigma \sqrt{1 - r_{12}}$$

A probable error of measurement of 5 means that in fifty per cent of the cases the obtained score will differ by as much or more than 5 from the pupil's true score. In fifty per cent of the cases the difference will be less.

Prognostic test. A prognostic test is a test which has for its function the prediction of a pupil's status at some future time. This prediction, of course, is based upon the pupil's performance at the present time. All tests have some prognostic value, but certain tests which have been devised with special reference to this function are called prognostic tests. (See pages 307 and 321.)

Quality. The quality of a pupil's performance is sometimes described in terms of the per cent of the exercises which he has done correctly. In such cases quality is synonymous with accuracy. Certain types of performances (for example, a specimen of handwriting) cannot be classified as right or wrong. In such cases quality means merit and it is described in terms of a quality scale.

Quotient score. A point score or an age score is simply a description of the absolute amount of a pupil's achievement or general intelligence. Such absolute measures are significant only when compared with appropriate norms. For this reason it has been proposed to divide the point scores or age scores by certain other measures of the pupil. For example, a pupil's mental age divided by his chronological age gives a quotient which is called the intelligence quotient or I.Q. A pupil's achievement age divided by his mental age gives the achievement quotient or A.Q. More strictly speaking, the A.Q. is the quotient of a pupil's achievement age divided by the norm for his mental age. Other quotients have been proposed. For example, a pupil's achievement age divided by his chronological age gives the educational quotient or E.Q. The educational quotient divided by the intelligence quotient has been called the accomplishment quotient or A.Q. This, however, is identical with the achievement quotient described above. (See page 380.)

Rate score. A rate score is a measure of a pupil's rate of work. It is usually expressed in terms of the number of exercises or the number of units of work which he has attempted within a given time limit. It may, however, be expressed as the number of minutes or seconds used by a pupil to complete a specified amount of work.

Rate test. A rate test is one which yields a rate score. It may yield other scores also but it is essential that it yields a rate score which is unaffected by the other dimensions of the pupil's performance.

Recognition exercise. In a recognition exercise the pupil is given a list of words or statements and asked to check the one which is the correct answer to the question asked. Test-makers have devised a number of ingenious modifications of this simple form of the recognition exercise. (See pages 121 and 290.)

Reliability. The reliability of a test describes the extent to which a second application of a test will yield scores equivalent to the first. It is a well-known fact that when a test is administered the second time, some pupils will make higher scores and some lower. These changes are due, for the most part, to the presence of variable errors in both sets of scores. The reliability of a test is the description of the magnitude of these variable errors. Any constant errors produced by practice effect or by inaccurate timing or by other conditions which affect the entire group are not included in the reliability. (See Coefficient of reliability, Index of reliability, Probable error of estimate, and Probable error of measurement.)

Scale. When used in a restricted sense the word "scale" designates that portion of a measuring instrument which is used in describing a pupil's performance. In the case of some of our measuring instruments the scale is conspicuous, as, for example, in Willing's Scale for Measuring Written Composition. This scale is used only in describing the performance of pupils. In order to secure a suitable performance it is necessary to follow certain directions which are not, strictly speaking, a part of this scale. In other measuring instruments, such as Courtis Standard Research Tests in Arithmetic, Series B, the scale is less obvious. There is, however, in every measuring instrument a scale which functions in the description of the performances secured from the pupils. The word "scale" is used also in a general sense to designate the total measuring instrument. Usually this is done only when the scale for describing the pupil's performance is the distinguishing characteristic of the measuring instrument. (See Test.)

Scaled test. A scaled test is one in which the exercises are arranged in order of ascending difficulty. Usually, the increase in difficulty from one exercise to the next is approximately constant throughout the scale. This is a desirable but not necessary feature. Another essential characteristic of the scaled test is that the exercises of least difficulty be sufficiently easy so that all pupils to whom the test is given will be able to do them and that the most difficult exercises be such that practically no pupils will be able to do them correctly.

School mark. This phrase is synonymous with "grade." Sometimes the word "mark" is used alone. (See Grade.)

Score. A pupil's score is a description of his performance. There are several types of scores, each of which has its own function. (See Rate score, Accuracy, Quality, Difficulty, Point score, Derived score, Combined dimensions.)

Selection of exercises. Usually in constructing educational tests a large number of exercises are secured and from this collection those to be used in the final test are selected. There are three criteria of selection which are frequently used, sometimes singly and sometimes in combination: (1) statistical selection, (2) agreement with educational objectives, and (3) suitability for testing purposes as determined by trial. Occasionally the selection is made by the author of the test without the guidance of definite criteria. Such selection may be described as arbitrary. (See Statistical selection and Educational objectives.)

Sigma (σ). (See Standard deviation.)

Spiral test. The word "spiral" has been used to describe a measuring instrument which consists of several sub-tests so arranged that in general there is an increase in difficulty in the successive sub-tests. A good example of this type of test is the Cleveland-Survey Arithmetic Test.

Standard deviation. The standard deviation is a measure of the spread or variability of a distribution. It tells how widely the measures or scores are spread out from the average. Two abbreviations are used for the standard deviation, S.D. and σ (sigma). If we form an interval extending from -1.0σ (i.e., 1.0σ less than the average) to $+1.0\sigma$, there will be included 68.26 per cent of the total distribution. (See Probable error.)

Standard distribution. The distribution of scores or grades which has been accepted as standard is called a "standard distribution." This may or may not coincide with the normal distribution. (See page 429.)

Standardized test. A test is said to be standardized when norms or standards have been determined for it. The standardization of the test has no reference to the selection of the exercises or to the unit in terms of which the point score is expressed. In the field of physical measurement, the standardization of a measuring instrument has a different meaning. It refers to the fixing of the magnitude of the unit. For example, the standardization of linear measures means fixing the precise length of the fundamental unit — the yard. This meaning of standardization is approached in some of the proposed derived scores.

Standards. (See Norms.)

Statistical selection of exercises. The usual procedure in constructing an educational test is to secure a rather large collection of exercises. From this list certain exercises are selected. One method for making this selection is to ascertain the per cent of correct responses for each exercise and from this to compute their difficulty. Those exercises are then selected whose degree of difficulty is appropriate for the structure of the desired test. Such a selection is said to be statistical. (See Educational objectives.)

Subject age. A pupil's subject age is his achievement age in a given subject, such as arithmetic, reading, etc. (See Achievement age.)

Subjective. An educational test is said to be subjective when different persons or the same person at different times, using it to measure the same thing, secure different results. The source of the subjectivity may be in the giving of the tests to the pupils or in the scoring of the test papers. In the latter case the scoring or the description of the pupil's performance is said to be subjective. This means that different persons will tend to assign different scores to the same papers. It should be noted that "subjective" and "objective" are relative terms. All educational tests are subjective in some degree. Certain tests are very highly subjective and others are only very slightly so. As the term is generally used, a subjective test is one which is highly subjective. (See Objective.)

Subject quotient. A pupil's subject quotient is merely his educational quotient or achievement quotient for a particular subject. (See Quotient score.)

Sub-test. Some measuring instruments consist of major divisions which are called sub-tests. For example, the Cleveland-Survey Test in Arithmetic is a measuring instrument which consists of fifteen sub-tests. Each sub-test is made up of a number of exercises. (See Exercise.)

Survey test. A survey test is one which is general in its scope. It is usually made up of a number of sub-tests covering a variety of fields of subject-matter. The scores yielded by these sub-tests may or may not be combined into a single score. The function of a survey test is to yield a general or average measure of a pupil's achievement over a large field. Sometimes this field may be restricted to certain divisions within a subject as, for example, arithmetic, or it may include several school subjects.

Teachers' marks. The term "teachers' marks" is frequently used with the same meaning as "grades." Since the word "grade" is used also to designate a pupil's classification in a school system, the use of another term to designate the words and symbols used to express a pupil's standing has certain advantages. However, the word "grade" is widely used for this purpose.

Test. The word "test" is used both in a general sense and in a restricted sense. In the general sense it is used to designate any type of instrument for measuring mental ability. Thus it may be used in referring both to instruments which have been named "tests" and to instruments which have been named "scales" by their authors. In the restricted sense it refers to the portion of a measuring instrument that is used to secure a performance from the pupil. Some of our measuring instruments are spoken of as tests and others as scales, but there is little evidence of discrimination in the use of these terms. In so far as there has been discrimination in respect to "test" and "scale," that term has been used which was most characteristic of the distinguishing fea-

ture of the measuring instrument. For example, we have the Courtis Arithmetic Tests, the Kansas Silent Reading Test, and the Thorndike Handwriting Scale. (See Scale, Uniform test, Scaled test, Irregular test, Cycle test, and Spiral test.)

Time limit. A test is said to be "timed" when the time allowed is such that a measure of the rate of work of the pupils can be secured. Usually this means that the time limit is such that practically no pupils will be able to finish the test. All types of test may be timed, but the time limit is most significant in the case of a uniform test. When applied to a scaled test, if the time limit is such that practically all pupils are able to advance as far along the scale as their ability permits before time is called, the test is essentially untimed. Although a time limit may be specified in such a case, it is not incorrect to say that the pupils are allowed practically unlimited time or all the time they need.

True-false test. This is a test in which the pupil is asked to indicate whether statements are true or false. (See page 481.)

True score. A pupil's true score is defined as the average of a large number of measurements of a given ability made under the same conditions. It is, of course, impossible to make even a second measurement of a pupil's ability under exactly the same conditions as the first measurement was made because the taking of the test in itself has changed one factor of the testing conditions. For this reason it is impossible to obtain a true score by averaging the scores obtained from the repeated applications of a test. However, the concept of a true score is frequently helpful, and we are able to make certain statistical calculations with reference to true scores even though it is impossible to obtain them. (See Index of reliability and Probable error of measurement.)

T-score. The T-score is a derived score proposed by McCall. (See page 124.)

Uniform test. A uniform test is one whose exercises are approximately equivalent in difficulty. Generally the exercises are also similar in content. This equivalence in difficulty may be secured by constructing exercises of the same sort as, for example, in the Courtis Standard Research Tests in Arithmetic, Series B, or by selection on a statistical basis.

Validity. The term "validity" refers to the truthfulness with which a test fulfills its function. A test may fail to do this by reason of inaccurate scores or by failing to measure the ability specified by its function. A test whose score is lacking in accuracy is said to be unreliable. Such a test can never be highly valid. Because we are not able to obtain completely valid measures for purposes of comparison, it is necessary to use certain indirect and partial methods in determining the validity of a given test. (See Subjective, Reliability, and Discrimination.)

Variable errors. Variable errors are different for the different mem-

bers of a group. Some are zero, nearly half of them are positive and an equal number are negative. The distinguishing characteristic of all variable errors is this difference from pupil to pupil. Unless highly accurate measures of the same trait are available for comparison we are not able to determine the magnitude of the variable error for a particular pupil. The best we are able to do is to state what the chances are that the variable error does not exceed a certain magnitude in a particular case. This is done by using the Probable error of measurement. (See Reliability and Probable error of measurement.)

Vocational guidance. Vocational guidance relates to advising students with reference to their preparation for a vocation and also with reference to the vocation for which they are preparing. It overlaps with educational guidance. In fact, educational guidance may be considered vocational guidance when it is given with explicit reference to preparation for a specific vocation. (See page 456.)

APPENDIX B

LIST OF STANDARDIZED TESTS DESCRIBED

THE standardized tests described in Chapters II to IX are listed here, with the name of the publisher and the price at the time of going to press. The page reference in the margin is to the description of the test in the body of the text. Except when otherwise indicated, the prices are for one hundred copies of the test and four sets of directions and other accessories.

An attempt has been made to secure accurate prices, but one should remember that most publishers reserve the right to change prices without notice. For this reason too much dependence should not be placed upon the prices given. They will, however, suffice to guide one in taking account of the cost of test materials in making an intelligent selection from the tests listed. Some of the publishers allow discount when the tests are purchased in quantity. In practically all cases the purchaser is required to pay transportation charges in addition to the prices quoted.

ARITHMETIC

- | | |
|--|------|
| Buckingham Scale for Problems in Arithmetic | 64 |
| Division 1, Grades III and IV | |
| Division 2, Grades V and VI | |
| Division 3, Grades VII and VIII | |
| Forms 1 and 2 | |
| Public School Publishing Company, Bloomington, Illinois. | 80c. |
| Cleveland-Survey Tests in Arithmetic | 36 |
| Grades III to VIII | |
| Public School Publishing Company, Bloomington, Illinois. | |
| \$1.90. | |
| Courtis Standard Research Tests, Series B | 27 |
| Grades IV to VIII | |
| Forms 1, 2, 3, and 4 | |
| S. A. Courtis, 1807 East Grand Boulevard, Detroit, Michigan. | |
| \$1.72. | |

- Courtis Supervisory Tests in Arithmetic** 34
 Test A and Test B, Grades IV to VIII
 Forms 1, 2, 3, and 4
 S. A. Courtis, 1807 East Grand Boulevard, Detroit, Michigan.
 66c. each test; Manual of directions, 20c. extra.
- Lunceford Diagnostic Tests in Addition** 40
 Primary Grades
 Forms 1 and 2
 Bureau of Educational Measurements and Standards, Kansas
 State Normal School, Emporia, Kansas. 75c.
- Monroe Diagnostic Tests in Arithmetic** 41
 Part I, Integers, Grades IV to VIII
 Part II, Integers, Grades IV to VIII
 Part III, Common Fractions, Grades V to VIII
 Part IV, Decimal Fractions, Grades VI to VIII
 Public School Publishing Company, Bloomington, Illinois. 85c.
 for each part.
- Monroe General Survey Scales in Arithmetic** 45
 Scale 1, Grades III, IV, and V
 Scale 2, Grades VI, VII, and VIII
 Forms 1, 2, and 3
 Public School Publishing Company, Bloomington, Illinois. \$1.
- Monroe Standardized Reasoning Tests in Arithmetic** 61
 Test 1, Grades IV and V
 Test 2, Grades VI and VII
 Test 3, Grade VIII
 Forms 1 and 2
 Public School Publishing Company, Bloomington, Illinois. 80c.
- Peet-Dearborn Progress Tests in Arithmetic** 57
 Intermediate series, Grades IV and V
 Upper-grade series, Grades VI, VII, and VIII
 Houghton Mifflin Company, Boston, New York, Chicago, and
 San Francisco. \$4.80.
- Stone Reasoning Test** 66
 Grades V to VIII
 Bureau of Publications, Teachers College, Columbia Univer-
 sity, New York City. 40c.; Manual of directions, 65c. extra.
- Woody Arithmetic Scales** 49
 Grades III to VIII
 Series B is an abbreviated form of Series A
 A second form of these scales has been prepared by W. W.

Theisen and published by the Parker Company, Madison, Wisconsin.

Bureau of Publications, Teachers College, Columbia University, New York City. Series A, 50c. each scale; Series B, \$1.50; Manual of directions, 60c. extra.

- Woody-McCall Mixed Fundamentals 56
 Grades II to VIII
 Forms 1 and 2

Bureau of Publications, Teachers College, Columbia University, New York City. 60c.

BATTERIES OF EDUCATIONAL TESTS

- Illinois Examination (Illinois General Intelligence Scale, Monroe's Standardized Silent Reading Tests, Monroe's General Survey Scale in Arithmetic) 382
 Examination I, Grades III, IV, and V
 Examination II, Grades VI, VII, and VIII
 Public School Publishing Company, Bloomington, Illinois. \$4.

- Lippincott-Chapman Classroom Products Survey Tests (Arithmetic fundamentals, arithmetic problems, reading continuous passage, reading selections) 386
 Grades V to VIII
 J. B. Lippincott Company, 227 South Sixth Street, Philadelphia. \$3.50.

- Pintner Educational Survey Tests (Arithmetic, reading, completion grammar, geography, and history) 383
 Grades II to VIII
 Some of the sub-tests in this battery are abbreviated forms of well-known tests. Pintner has also devised a battery of non-language mental tests which are designed to be used in connection with these educational survey tests.
 College Book Store, Columbus, Ohio. \$8.

- Pressey Second Grade Attainment Scale (Reading, arithmetic, spelling) 387
 Forms 1 and 2
 Public School Publishing Company, Bloomington, Illinois. \$1.

- Pressey Third Grade Attainment Scale (Spelling, arithmetic, and silent reading) 387
 Forms 1 and 2
 Public School Publishing Company, Bloomington, Illinois. \$1.50.

- Pressey Scale of Attainment No. 2 (History, arithmetic, and English) . 388
 Grade VIII
 Department of Psychology, Indiana University, Bloomington,
 Indiana. \$1.65.
- Stanford Achievement Test 384
 Primary Examination, Grades II and III
 Advanced Examination, Grades IV to VIII
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. Primary Examination, \$5.60; Ad-
 vanced Examination, \$8.00; Manual of directions, 30c. extra.

ENGLISH

(Under the head of English we have included measuring instruments for a variety of subdivisions of the general field. Composition scales, tests in language, grammar, punctuation, and literature are to be found under this head. Spelling scales are listed separately.)

- Abbott-Trabue Exercises for Judging Poetry 265
 Series X and Series Y
 These series are to be used as duplicate forms.
 Bureau of Publications, Teachers College, Columbia Univer-
 sity, New York City. \$7.50; Bulletin of directions, 40c.
- Briggs English Form Test 245
 Grades VII and VIII and High School
 Forms Alpha and Beta
 Bureau of Publications, Teachers College, Columbia Univer-
 sity, New York City. \$1.40.
- Charters Diagnostic Language Tests 242
 Pronouns, Verbs, Miscellaneous A and Miscellaneous B
 Grades III to VIII
 Forms 1 and 2
 Public School Publishing Company, Bloomington, Illinois. 80c.
- Charters Diagnostic Language and Grammar Tests 245
 Pronouns, Verbs, Miscellaneous
 Grades VII and VIII
 Public School Publishing Company, Bloomington, Illinois.
 \$1.50.
- Courtis Standard Research Test in Composition 259
 S. A. Courtis, 1807 East Grand Boulevard, Detroit, Michigan.
 25c. per set.

- Hillegas Scale for Measurement of English Composition by Young People** 256
 Grades IV to XII
 The Hudelson Scale and the Nassau County Supplement to the Hillegas Scale are essentially revisions of this scale. In general they will be found more satisfactory than the original scale. The Thorndike Extension of the Hillegas Scale is another revision which has corrected some of the faults of the original scale.
 Bureau of Publications, Teachers College, Columbia University, New York City. 3c. per copy.
- Hudelson English Composition Scale** 258
 Grades IV to XII
 World Book Company, Yonkers, New York, and 2126 Prairie Avenue, Chicago, Illinois. 56c. per copy.
- Hudelson Typical Composition Ability Scale** 258
 Grades IV to XII
 Public School Publishing Company, Bloomington, Illinois. 10c. per single copy; Teacher's Handbook, 20c. per copy.
- Kirby Grammar Test** 246
 Grades VII to XII
 Extension Division, University of Iowa, Iowa City, Iowa. \$1.75.
- Lewis English Composition Scales** 259
 Grades V to XII
 World Book Company, Yonkers, New York, and 2126 Prairie Avenue, Chicago, Illinois. 25c. each scale.
- Nassau County Supplement to the Hillegas Scale** 257
 Bureau of Publications, Teachers College, Columbia University, New York City. 10c. per copy.
- Pressey Diagnostic Tests in English Composition (Vocabulary, grammar, and punctuation)** 247
 Revision of original tests now published by Public School Publishing Company, Bloomington, Illinois.
- Starch Punctuation Scale** 249
 Public School Publishing Company, Bloomington, Illinois. 80c.
- Van Wagenen English Composition Scales** 259
 Grades V to XII
 World Book Company, Yonkers, New York, and 2126 Prairie Avenue, Chicago, Illinois. 25c. per copy for each scale.
- Van Wagenen Reading Scale for English Literature** 265
 Forms A, B, and C
 Public School Publishing Company, Bloomington, Illinois. \$3.

- Willing Scale for Measuring Written Composition 258
 Public School Publishing Company, Bloomington, Illinois. 9c.
 per copy.
- Wilson Language Error Test 249
 Grades III to XII
 Forms A, B, and C
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$5.00.

GEOGRAPHY

- Courtis Standardized Supervisory Tests in Geography 275
 States and important cities in United States
 The world — oceans, continents, and countries
 Grades III to VIII
 Forms A and B
 S. A. Courtis, 1807 East Grand Boulevard, Detroit, Michigan.
 \$2.68.
- Gregory-Spencer Geography Tests 277
 Grades VI, VII, and VIII
 Forms A, B, and C
 Bureau of Educational Research, University of Oregon, Eugene,
 Oregon. \$4.
- Hahn-Lackey Geography Scale 274
 Grades IV to VIII
 H. H. Hahn, State Normal School, Wayne, Nebraska. 16c.
 single copy.
- Posey-Van Wagenen Geography Scales 279
 Information R and Thought S
 Division I, Grades V and VI
 Division II, Grades VII and VIII
 Public School Publishing Company, Bloomington, Illinois. \$1.50.
- Witham Standard Geography Tests 281
 Test 1 — the World
 Test 2 — United States
 Test 3 — South America
 Test 4 — Europe
 Test 5 — Asia
 Test 6 — Africa
 Test 7 — North America
 Test 8 — Commerical Geography
 Used in grades in which these topics are taught.
 J. L. Hammett Company, Cambridge, Massachusetts. \$3.50
 for each test.

HANDWRITING

- Ayres Measuring Scale for Handwriting (three-slant edition)** 166
 Grades III to VIII
 Russell Sage Foundation, Division of Education, New York
 City. 18c. single copy.
- Ayres Scale, "Gettysburg Edition"** 166
 Grades II to VIII
 Russell Sage Foundation, Division of Education, New York
 City. 10c. single copy.
- Freeman Chart for Diagnostic Faults in Handwriting** 170
 Grades II to VIII
 Houghton Mifflin Company, Boston, New York, Chicago, and
 San Francisco. 30c.
- Gray Standard Individual Score Card for Measuring Handwriting** . 172
 Grades II to VIII
 Form II for use of individual pupils
 Form III for use as wall chart for class use
 Public School Publishing Company, Bloomington, Illinois.
 Form II, 75c. per hundred; Form III, 10c. single copy.
- New York City Penmanship Scale** 170
 Grades II to VIII
 The Macmillan Company, New York City. 25c. per copy.
- Starch Handwriting Scale, Revised** 169
 Grades I to VIII
 University Coöperative Company, 504 State Street, Madison,
 Wisconsin. 50c. single copy.
- Thorndike Handwriting Scale** 164
 Grades V to VIII
 Bureau of Publications, Teachers College, Columbia University,
 New York City. 12c. single copy.

HISTORY

- Barr Diagnostic Tests in American History** 290
 Primarily for use in high schools
 Series A and Series B
 Public School Publishing Company, Bloomington, Illinois. \$4.
- Gregory Tests in American History** 293
 Form A
 Bureau of Educational Research, University of Oregon, Eugene,
 Oregon. \$4.

- Hahn Scale for Measuring Ability of Children in History** 283
 Grades VII and VIII
 H. H. Hahn, State Normal School, Wayne, Nebraska. 16c.
 single copy.
- Harlan Test of Information in American History** 284
 Grades VII and VIII
 Public School Publishing Company, Bloomington, Illinois. 80c.
- Van Wagenen American History Scales** 286
 Information Scale A and Scale B
 Thought Scale A and Scale B
 Character Judgment Scale A and Scale B
 Bureau of Publications, Teachers College, Columbia University,
 New York City. \$1.25 each scale; Manual, 96c. extra.

INTELLIGENCE TESTS

INDIVIDUAL INTELLIGENCE TESTS

For individual testing the Stanford Revision of the Binet-Simon Tests is the most widely used. It is published by Houghton Mifflin Company, Boston, New York, Chicago, and San Francisco. Recently an extension of the Binet-Simon Scale by F. Kuhlmann has been published by Warwick and York, Baltimore, Maryland. The World Book Company, Yonkers, New York, and Chicago, Illinois, has recently published the Herring Revision of the Binet-Simon Tests. This revision has received some very favorable comment.

GROUP INTELLIGENCE TESTS

- Army Group Examination, Alpha** 353
 For use in high schools and colleges
 Bureau of Educational Measurements and Standards, Kansas
 State Normal School, Emporia, Kansas, \$3; Manual of directions,
 75c.; Stencils, \$1.25.
- Dearborn Group Test of Intelligence, Series I, Revised Edition** . . . 366
 Grades I to III
 J. B. Lippincott Company, 227 South Sixth Street, Philadelphia.
 \$4.50.
- Dearborn Group Test of Intelligence, Series II, Revised Edition** . . . 366
 Grades IV to IX
 This series of general intelligence tests consists of two parts —
 General Examination C and General Examination D. They are
 non-verbal in character.
 J. B. Lippincott Company, 227 South Sixth Street, Philadel-
 phia, Pennsylvania. \$4.50.

LIST OF TESTS DESCRIBED

511

- Detroit First-Grade Intelligence Test** 366
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$5.80.
- Haggerty Intelligence Examinations** 355
 Delta I, Grades I to III
 Delta II, Grades III to IX
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. Delta I, \$5.60; Delta II, \$5.32; Man-
 ual of directions, 25c.
- Illinois General Intelligence Scale** 355
 Grades III to VIII
 Public School Publishing Company, Bloomington, Illinois. \$2.
- Myers Mental Measure** 366
 Grades I to XII
 Forms 1 and 2
 Newson and Company, 73 Fifth Avenue, New York City. \$5.
- National Intelligence Tests** 355
 Scale A and Scale B
 Forms 1 and 2 of each scale
 Grades III to VIII
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$5.20 for each scale; Manual of di-
 rections, 20c.
- Otis Group Intelligence Scale, Primary Examination** 355
 Forms A and B
 Grades I to IV
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$5.00; Manual of directions, 30c.
- Otis Group Intelligence Scale, Advanced Examination** 355
 Forms A and B
 Grades VII to XII
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$5.20; Manual of directions, 30c.
- Pressey Primary Classification Test** 366
 Grades I and II
 This is a revision of the original Pressey Primer Scale which has
 been widely used.
 Public School Publishing Company, Bloomington, Illinois.
 \$1.50.

- Pressey Intermediate Classification Test 361
 Grades III to VI
 Forms A and B
 Public School Publishing Company, Bloomington, Illinois.
 \$1.25.
- Pressey Senior Classification Test 361
 Grades VII and VIII
 Forms A and B
 Public School Publishing Company, Bloomington, Illinois.
 \$1.25.
- Terman Group Test of Mental Ability 315
 Forms A and B
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$5.40.

LATIN

- Godsey Diagnostic Latin Composition Test 314
 Mason D. Gray, East High School, Rochester, New York. \$1.
- Henmon Latin Tests 315
 Tests 1, 2, 3, and 4
 These tests are to be considered as duplicate forms. Each consists of two parts — vocabulary and sentences. Test X is limited to vocabulary.
 World Book Company, Yonkers, New York, and 2126 Prairie Avenue, Chicago, Illinois. \$2.
- Holtz-Godsey Latin Teaching Tests 316
 Bureau of Educational Measurements and Standards, Kansas
 State Normal School, Emporia, Kansas. 50c.
- Pressey Test in Latin Syntax 317
 Public School Publishing Company, Bloomington, Illinois. \$2.
- Starch-Waters Latin Tests 317
 University Coöperative Company, 504 State Street, Madison, Wisconsin. \$2.
- Tyler-Pressey Test in Latin Verb Forms 317
 Mason D. Gray, East High School, Rochester, New York. \$1.
- Ullman-Kirby Latin Comprehension Test 318
 Mason D. Gray, East High School, Rochester, New York. \$1.

MATHEMATICS

- Douglass Standard Diagnostic Tests for First-Year Algebra** 303
 Series A (for the fundamental operations)
 Series B (for additional processes)
 Bureau of Educational Research, University of Oregon, Eugene,
 Oregon. Series A, \$1.60; Series B, \$3.50.
- Hotz Algebra Scales** 304
 Test 1, Addition and Subtraction
 Test 2, Multiplication and Division
 Test 3, Equation and Formulation
 Test 4, Graphs
 Test 5, Problems
 Series A and B. The former is an abbreviated form of the latter.
 Bureau of Publications, Teachers College, Columbia University,
 New York City. \$1.25; Manual of directions, 75c.
- Illinois Standardized Algebra Tests** 305
 Public School Publishing Company, Bloomington, Illinois.
 \$2.50.
- Minnick Geometry Tests** 309
 Test A, Drawing of figures
 Test B, Stating of hypotheses and conclusions
 Test C, Recalling facts and figures
 Test D, Selecting and organizing facts to produce a proof
 Public School Publishing Company, Bloomington, Illinois.
 \$2.50 per each test.
- Rogers Tests of Mathematical Ability** 307
 Bureau of Publications, Teachers College, Columbia University,
 New York City. \$9; Manual of directions, 65c.
- Rugg and Clark Tests in First-Year Algebra** 306
 University of Chicago Book Store, University of Chicago,
 Chicago, Illinois. \$8.

MODERN LANGUAGES

- Handschin Comprehension and Grammar Test A: French** 320
 This test is designed for use in the first year.
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$2.
- Handschin Silent Reading Test A: French** 320
 This test is designed for either first or second year classes.
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$2.

- Handschin Silent Reading Test B: French** 320
 This test is designed for use in either the first or second year.
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$2.
- Handschin Silent Reading Test A: Spanish** 319
 This test is designed for either first or second year classes.
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$2.
- Handschin Silent Reading Test B: Spanish** 320
 This test is designed for use in either the first or second year.
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$2.
- Henmon French Tests** 321
 Tests 1, 2, 3, and 4 constitute four duplicate forms.
 Each test consists of two parts — vocabulary and sentences.
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$2.
- Wilkins Prognosis Test in Modern Languages** 321
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$6.40.

READING

- Burgess Picture Supplement Scale for Measuring Ability in Silent Reading** 112
 Grades II to VIII
 Forms 1, 2, 3, and 4
 Division of Education, Russell Sage Foundation, New York.
 \$1.25.
- Courtis Silent Reading Test No. 2** 106
 Grades II to VI
 Forms 1, 2, and 3
 S. A. Courtis, 1807 East Grand Boulevard, Detroit, Michigan.
 \$3.
- Gray Standardized Oral Reading Paragraphs** 135
 Grades I to VIII
 Public School Publishing Company, Bloomington, Illinois. \$1.
- Haggerty Achievement Examination in Reading, Sigma 1** 118
 Grades I to III
 World Book Company, Yonkers, New York, and 2126 Prairie
 Avenue, Chicago, Illinois. \$4.40; Manual of directions, 25c. extra.

- Haggerty Achievement Examination in Reading, Sigma 3** 118
 Grades VI to XII
 Forms A and B
 World Book Company, Yonkers, New York, and 2126 Prairie Avenue, Chicago, Illinois. \$5.20; Manual of directions, 25c. extra.
- Holley Sentence Vocabulary Scale** 130
 Grades II to XII
 Public School Publishing Company, Bloomington, Illinois. 80c.
- Monroe Standardized Silent Reading Tests** 99
 Test I, Grades III, IV, and V
 Test II, Grades VI, VII, and VIII
 Test III, High School
 Forms 1, 2, and 3
 There is no Form 3 of Test III. Tests I and II have been revised (see below) and the use of the revised forms is recommended.
 Public School Publishing Company, Bloomington, Illinois.
 Tests I and II, 80c.; Test III, \$1.
- Monroe Standardized Silent Reading Tests, Revised** 102
 Test I, Grades III, IV, and V
 Test II, Grades VI, VII, and VIII
 Forms 1, 2, and 3
 Public School Publishing Company, Bloomington, Illinois. 80c.
- Thorndike Scale Alpha 2. For Measuring the Understanding of Sentences** 118
 Part I, Grades III to V
 Part II, Grades VI to XII
 Bureau of Publications, Teachers College, Columbia University, New York City. \$1.70.
- Thorndike-McCall Reading Scale for the Understanding of Sentences** 118
 Grades II to VIII
 Forms 1, 2, 3, 4, 5, and 6
 Bureau of Publications, Teachers College, Columbia University, New York City. \$2.
- Thorndike Test of Word Knowledge** 130
 Forms A, B, C, and D
 Bureau of Publications, Teachers College, Columbia University, New York City. \$1.50.
- Thorndike Visual Vocabulary Scales** 130
 Scale A-2, Series X and Series Y
 Scale B, Series X and Series Y
 Grades III to VIII
 Bureau of Publications, Teachers College, Columbia University, New York City. \$1.50.

SCIENCE

- Downing Revised Range-of-Information Test in Science 323
 Elliot R. Downing, School of Education, University of Chicago,
 Chicago, Illinois. 40c.
- Iowa Physics Tests 325
 Series A, Mechanics
 Series B, Heat
 Series C, Electricity and Magnetism
 Forms 1 and 2
 Public School Publishing Company, Bloomington, Illinois.
 \$2 each test.
- Starch Physics Test 324
 University Coöperative Company, 504 State Street, Madison,
 Wisconsin. \$2.
- Van Wagenen Reading Scales for General Science 325
 Scale A and Scale B
 Public School Publishing Company, Bloomington, Illinois. \$3.

SPELLING

- Ayres Spelling Scale for Measuring Ability in Spelling 206
 Grades III to VIII
 Division of Education, Russell Sage Foundation, New York.
 10c. single copy.
- Iowa Spelling Scales 207
 Scale I, Grades II, III, and IV
 Scale II, Grades IV, V, and VI
 Scale III, Grades VI, VII, and VIII
 Extension Division, University of Iowa, Iowa City, Iowa.
- Buckingham Extension of Ayres Spelling Scales 206
 Grades III to VIII and High School
 Public School Publishing Company, Bloomington, Illinois. 14c.
 single copy.
- Courtis Standard Dictation Tests 221
 Grades II to VIII
 These tests are in the form of timed sentences. There are two
 tests for each half grade, an initial test, Form A, and a final test,
 Form B.
 S. A. Courtis, 1807 East Grand Boulevard, Detroit, Michigan.
 11c. single copy.

- Monroe Timed Sentence Spelling Tests 219**
Test 1, Grades III and IV
Test 2, Grades V and VI
Test 3, Grades VII and VIII
Public School Publishing Company, Bloomington, Illinois. 12c.
per set of three.
- Sixteen Spelling Scales for Secondary Schools (Seven S Spelling
Scales) 221**
Bureau of Publications, Teachers College, Columbia Univer-
sity, New York City. 40c.

INDEX

Titles of tests are not included in this index. They will be found in Appendix B. The reader should also consult Appendix A for technical terms. A few of them have been repeated in the index, but there has been no attempt to have it complete in this respect.

- Accuracy of class scores, 72; of individual scores, 69.
 Age scores, 380.
 Algebra, fundamental operations, 302; practice tests, 311; problem of measurement, 301.
 Almack, J. C., 190.
 Arithmetic, problem of measurement, 19.
 Arithmetical abilities, nature, 19; significant dimensions, 23.
 Ashbaugh, E. J., 188.
 Ayres, L. P., 188, 223.
- Bagley, W. C., 284, 449.
 Baldwin, Bird T., 344, 346.
 Ballou, F. W., 218.
 Breed, F. S., 181, 182, 447.
 Breslich, E. R., 447.
 Brownell, Baker, 264.
 Burgess, May Ayres, 128.
- Carney, C. S., 457.
 Classification of pupils, 436.
 Coefficient of reliability, 72, 406.
 Colloton, C., 345.
 Colvin, S. S., 333, 367, 461.
 Combining scores, 378.
 Composition scales, directions for using, 253, 260; reliability, 262.
 Constant errors, in examination scores, 476; in rate of handwriting, 180; in school marks, 3.
 Courtis, S. A., 21, 27, 34, 67, 70, 198, 213.
 Courtis Standard Practice Tests in Arithmetic, 87.
 Courtis Standard Practice Tests in Handwriting, 198.
 Courtis Practice Tests in Spelling, 235.
 Critical evaluation of a test, 403.
 Culp, Vernon, 182.
 Cutten, George B., 442.
 Cycle principle, 305.
- Dalman, Murray A., 311.
 Dalman hurdles in Algebra, 311.
 Derived scores, 124, 380.
- Diagnosis in handwriting, 176.
 Diagnosis, technique of, 74 ff.
 Diagnostic Record Sheet, 40, 53, 110, 227.
 Diagnostic tests, 26.
 Dickson, Virgil E., 340.
 Distribution of grades, 427.
 Doll, Edgar A., 345.
 Duplicate forms, construction, 47.
- Educational guidance, 453.
 Educational guidance policies, 458.
 Elliot, E. C., 5.
 English, problem of measurement, 240.
 Errors. *See* Constant errors and Variable errors.
 Errors, types of in Algebra, 312.
 Examination grades, reliability, 469.
 Examinations, rules for, 485.
 Examples, types of in Arithmetic, 21.
 Exercises, types of, 398; selection, 400.
- Freeman, F. N., 157, 158, 170, 178, 185, 197, 228, 345.
- Gates, Arthur I., 128, 135, 474.
 General intelligence, definition, 333; relation to school success, 460.
 General Tests, 26.
 Geography, problem of measurement, 273.
 Geyer, Denton L., 445, 446.
 Gilchrist, E. P., 174.
 Goddard, H. H., 346.
 Graves, F. P., 441.
 Graves, S. Monroe, 199.
 Gray, C. T., 144, 172, 183, 184.
 Gray, W. S., 143.
- Haggerty, M. E., 355, 364, 473.
 Handwriting, measurement of rate, 161; movement, 158; position, 157; problems of measurement, 155; reliability of measurement, 181.
 Handwriting scales, training in use of, 183; described, 164 ff.
 Henmon, V. A. C., 197, 366.
 History, problem of measurement, 282.

- Hudelson, Earl, 262, 264.
 Hurt, A. O., 183.
- Intelligence, definition, 332; relation to school success, 460.
 Intelligence quotient (I.Q.), 335, 340; constancy of, 343.
 Interpreting the scores of a class, 192.
- Johnson, F. W., 4.
 Johnson, Harry, 182.
 Jones, N. F., 228.
 Jordan, R. H., 345.
 Judd, C. H., 156, 158.
- Kallom, A. W., 20.
 Kelley, T. L., 174, 181, 407.
 Kelly, F. J., 99, 183, 263.
 King, Irving, 182.
 Knollin, H. E., 410.
 Koos, L. V., 189.
 Kuhlman, F., 145, 347.
- Lackey, E. E., 275.
 Latin, problem of measurement, 313.
 Lewis, E. E., 183, 188, 190.
 Lister, C. C., 170.
 Lull, H. G., 234.
- Manual, H. T., 182.
 Martens, Elise H., 340.
 McCall, W. A., 470, 474.
 Miller, W. S., 446, 461.
 Misspellings, types of, 229; causes of, 232.
 Modern languages, problem of measurement, 319.
 Morton, R. L., 182.
 Multiple track plan, 439.
 Myers, G. C., 366.
- National Research Council, 355.
 New examination, 480.
 Nifenecker, E. A., 170, 190.
 Norms, Courtis Standard Research Tests, Series B., 33; Courtis Supervisory Tests in Arithmetic, 35; Cleveland-Survey Arithmetic Tests, 39; Monroe's Diagnostic Tests in Arithmetic, 44; Monroe's General Survey Scales, in Arithmetic, 49; Woody Arithmetic Scales, 56; Monroe's Standardized Reasoning Tests in Arithmetic, 63; Buckingham's Scale for Problems in Arithmetic, 65; Stone's Reasoning Tests, 67; Monroe's Standardized Silent Reading Tests, Revised, 105; Courtis Silent Reading Test, No. 2, 112; Burgess Picture Supplement Scale, 115; Haggerty Reading Examinations, 126; Thorndike-McCall Reading Scale, 126; Gray's Oral Reading Test, 143; Handwriting, norms of progress, 185; Norms of attainment, 188; Charters's Diagnostic Language Test, 244; Starch Grammatical Scales, 248; National Standards for English Composition, 262.
 Nutt, H. W., 158, 199.
- O'Brien, F. J., 456.
 Odell, Charles W., 444.
 Otis, A. S., 211, 213, 219, 352, 363, 410.
- Paterson, Donald, 349.
 Pintner, R., 181, 349.
 Point scores, 380.
 Practice Tests, in Arithmetic, 87; in Algebra, 311; in Handwriting, 198; in Spelling, 235.
 Problem of measurement, 301. *See also* Arithmetic, Algebra, and other school subjects.
 Problem solving process, 58.
 Proctor, W. M., 461.
 Prognostic Tests, 299, 307, 321.
 Promotion and classification of pupils, 436.
- Quotient scores, 381.
- Reasoning, 59.
 Reliability, coefficient, 72, 406; of standardized tests, 470; of examinations, 470.
 Remedial instruction, in Arithmetic, 83; in Reading, 145 ff; in Handwriting, 198.
 Rhythm in Handwriting, 161.
 Robinson, Eleanor, 455.
 Rugg, H. O., 284, 345.
- Sackett, L. W., 163, 172, 182.
 School marks, 2.
 Science, problems of measurement, 323.
 Scientific management, 73.
 Scores, accuracy, 69; translation, 418, 425.
 Segregation, effect, 448, 450.
 Silent Reading tests, general limitations of, 98.
 Silent reading, nature of, 94; significant dimensions of, 95; problem of measurement, 96.
 Smith, James H., 86.
 Sorting method, 175.
 Spelling Demons, 228.
 Spelling problems of measurement, 205.
 Standard distribution of grades, 429.
 Standardized Tests, 12; reliability of, 470.
 Standardization of a test, 401.
 Standards, 419, 420.
 Standards. *See* Norms.
 Standards of accomplishment, 424.
 Starch, Daniel, 5, 67, 166, 222, 247.
 Stecher, L. I., 344, 346.

- Stenquist, J. L., 364, 366.
Stone, C. R., 45.
Stone, C. W., 20, 66.
Streitz, Ruth, 366.
Studebaker economy practice exercises, 88.
Subjective norms, 7.
- T-score, 124, 381.
Terman, L. M., 337, 341, 343, 344, 345, 346, 355.
Terman's five-track plan, 438.
Test construction, principles, 394.
Testing program, planning, 377, 388-89.
Tests, critical evaluation of, 403; types of, 395.
Theisen, W. W., 50.
Thorndike, E. L., 67, 164, 183, 211, 263, 355, 454-55.
- Thurston, L. L., 353.
Trabue, M. R., 263, 460.
- Variable errors, 7, 69; in examination scores, 469.
Vocational guidance, 456.
- Whipple, G. M., 335, 349.
Whitcomb, M. Edith, 342, 355.
Witham, Ernest, 172.
Wood, Ben D., 6.
Woody, Clifford, 52.
Written examinations (*see* Examinations), evidence of errors, 5 ff.
- Yerkes, Robert M., 350, 355, 462.

RIVERSIDE TEXTBOOKS IN EDUCATION

Edited by ELLWOOD P. CUBBERLEY

Dean of the School of Education, Leland Stanford Junior University

History of Education

CUBBERLEY: The History of Education

CUBBERLEY: Readings in the History of Education

CUBBERLEY: A Brief History of Education

CUBBERLEY: Public Education in the United States

General Educational Theory

ALMACK AND LANG: Problems of the Teaching Profession

CHAPMAN AND COUNTS: Principles of Education

CUBBERLEY: An Introduction to the Study of Education

CUBBERLEY: Rural Life and Education

DOUGLASS: Secondary Education

GESELL: The Pre-School Child

INGLIS: Principles of Secondary Education

MCCRACKEN AND LAMB: Occupational Information in the Elementary School

PROCTOR: Educational and Vocational Guidance

SMITH: An Introduction to Educational Sociology

SNEDDEN: Problems of Secondary Education

THOMAS: Principles and Technique of Teaching

WALLIN: The Education of Handicapped Children

Methods

ALMACK: Education for Citizenship

BOLENIUS: Teaching Literature in the Grammar Grades and High School

DOUGLASS: Modern Methods of High School Teaching

FREELAND, ADAMS, HALL: Teaching in the Intermediate Grades

KENDALL AND MIRICK: How to Teach the Fundamental Subjects

KENDALL AND MIRICK: How to Teach the Special Subjects

MARTZ AND KINNEMAN: Social Science for Teachers

HOUGHTON MIFFLIN COMPANY

BOSTON NEW YORK CHICAGO DALLAS SAN FRANCISCO

RIVERSIDE TEXTBOOKS IN EDUCATION

Edited by ELLWOOD P. CUBBERLEY

Dean of the School of Education, Leland Stanford Junior University

Methods

MINOR: Principles of Teaching Practically Applied

NEWCOMB: Modern Methods of Teaching Arithmetic

STONE: Silent and Oral Reading

STORMZAND: Progressive Methods of Teaching

THOMAS: The Teaching of English in the Secondary School

THOMAS: Training for Effective Study

TRAFTON: The Teaching of Science in the Elementary School

WOOFER: Teaching in Rural Schools

Healthful Teaching and Healthful Schools

AVERILL: Educational Hygiene

AYRES, WILLIAMS, WOOD: Healthful Schools. How to Build, Equip, and Maintain Them

HOAG AND TERMAN: Health Work in the Schools

TERMAN: The Hygiene of the School Child

Administration and Supervision

ALMACK AND BURSCH: Administration of Consolidated and Village Schools

BRIGGS: The Junior High School

CUBBERLEY: The Principal and His School

CUBBERLEY: Public School Administration

CUBBERLEY: State School Administration

NUTT: The Supervision of Instruction

PERRY: Discipline as a School Problem

PITTENGER: An Introduction to Public School Finance

RUGG: Primer of Graphics and Statistics for Teachers

SEARS: Classroom Organization and Control

SEARS: The School Survey

HOUGHTON MIFFLIN COMPANY

BOSTON NEW YORK CHICAGO DALLAS SAN FRANCISCO

RIVERSIDE TEXTBOOKS IN EDUCATION

Edited by ELLWOOD P. CUBBERLEY

Dean of the School of Education, Leland Stanford Junior University

Administration and Supervision

SHOWALTER: A Handbook for Rural School Officers

WILLIAMS: Graphic Methods in Education

Psychology and Child Study

AVERILL: Elements of Educational Psychology

AVERILL: Psychology for Normal Schools

EDWARDS: Psychology of Elementary Education

FREEMAN: Experimental Education

FREEMAN: How Children Learn

FREEMAN: The Psychology of the Common Branches

PECHSTEIN AND MCGREGOR: Psychology of the Junior High School Pupil

PECHSTEIN AND JENKINS: Psychology of the Kindergarten-Primary Child

WADDLE: An Introduction to Child Psychology

WALLIN: Clinical and Abnormal Psychology

Educational Tests and Measurements

FREEMAN: Mental Tests

HINES: A Guide to Educational Measurements

MONROE: An Introduction to the Theory of Educational Measurements

MONROE: Measuring the Results of Teaching

MONROE, DE VOSS AND KELLY: Educational Tests and Measurements.
Revised and Enlarged Edition.

RUGG: Statistical Methods Applied to Education

TERMAN: The Intelligence of School Children

TERMAN: The Measurement of Intelligence

Test Material for use with The Measurement of Intelligence

Record Booklets. Sold only in packages of 25

Condensed Guide for the Binet-Simon Intelligence Tests

Abbreviated Filing Record Cards. 25 in package

HOUGHTON MIFFLIN COMPANY

BOSTON NEW YORK CHICAGO DALLAS SAN FRANCISCO